DOCUMENT RESUME

ED 366 642 TM 021 049

AUTHOR	Westers, Paul
TITLE	The Solution-Error Response-Error Model: A Method for the Examination of Test Item Bias.
REPORT NO	ISBN-90-9006674-8
PUB DATE	Dec 93
NOTE	134p.; Doctoral Dissertation, Twente University, The Netherlands.
PUB TYPE	Dissertations/Theses - Doctoral Dissertations (041)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	Ability; *Estimation (Mathematics); Ethnic Groups; Foreign Countries; Item Analysis; *Item Bias; Item Response Theory; Minority Groups; Models; Racial Differences; Sex Differences; Simulation; *Test Items; Test Use; Test Validity
IDENTIFIERS	Polytomous Items; Pseudo Likelihood Theory; *Rasch Model; *Solution Error Response Error Model

ABSTRACT

The subject of this dissertation is the examination of differential item functioning (DIF) through the use of loglinear Rasch models with latent classes. DIF refers to the probability that a correct response among equally able test takers is different for various racial, ethnic, and gender groups. Because usual methods of detecting DIF give little information about the reason an item is biased, use of the solution-error response-error (SERE) model of H. Kelderman is proposed. It is demonstrated that the SERE model can show whether DIF is caused by the difficulty of the item, the attractiveness of its alternatives, or both. The large amount of computer memory space required makes this method impractical for a large number of items. A new method is proposed based on the division of the whole item set into several subsets, which is made possible by the collapsibility of the SERE model. With the use of subsets of items, the parameters of the entire SERE model can be obtained only by simultaneous estimation of the parameters of the collapsed SERE models through use of pseudo-likelihood theory. A simulation study demonstrates that a distinction can be made between the two types of DIF using the new approach. A generalization of the SERE model applicable to polytomously scored latent states, that may be explained with a multidimensional latent space, is discussed. Five appendices illustrate applications of these models with reference to existing tests and the collapsed SERE model. (Contains 167 references.) (SLD)

****
*   Reproductions supplied by EDRS are the best that can be made   *
*   from the original document.   *
****

ED 366 642

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. NELISSEN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

THE SOLUTION-ERROR RESPONSE-ERROR MODEL: A METHOD FOR THE EXAMINATION OF TEST ITEM BIAS

P. Westers

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Westers, Paul

The solution-error response-error model: a method for the
examination of test item bias / Paul Westers. - [S.l. : s.n.]. - Ill.
Proefschrift Enschede. - Met lit. opg. - Met samenvatting in het Nederlands.
ISBN 90-9006674-8
Trefw.: testconstructie.

Druk: FEBODRUK te Enschede

THE SOLUTION-ERROR RESPONSE-ERROR MODEL: A METHOD FOR THE EXAMINATION OF TEST ITEM BIAS

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus
prof.dr. Th.J.A. Popma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 10 december 1993 te 13.15 uur

door

Paul Westers
geboren op 20 augustus 1960 te Groningen

Dit proefschrift is goedgekeurd door de promotor prof.dr. W.J. van der Linden.

Assistent-promotor: Dr. H. Kelderman

Promotiecommissie:
prof.dr. J.A.P. Hagenaars
prof.dr. G.J. Mellenbergh
prof.dr. W. Schaafsma
prof.dr. N.D. Verhelst
dr. M.P.F. Berger

To my parents and sister Simona

Voor mijn ouders en zus Simona

PREFACE

In this dissertation the results are discussed of the research project titled 'Item bias detection using the loglinear Rasch model with latent classes', which, under the auspices of the Interuniversity Graduate School of Psychometrics and Sociometrics (IOPS), was carried out at the Department of Educational Measurement and Data Analysis (OMD) of the Faculty of Educational Science and Technology of the University of Twente. Doing the research and writing this thesis would not have been possible without the support of a large number of people. Here I would like to thank everyone who has contributed to the realization of the dissertation. I owe special thanks to:

Wim J. van der Linden for his valuable advice and critical comments; Henk Kelderman for his expert supervision and the interesting discussions; The IOPS for the courses that enabled me to widen and/or deepen my knowledge of psychometrics and sociometrics; the OMD department for the stimulating and heartwarming atmosphere at work; Norman D. Verhelst, Jeroen Vermunt and Frank J. van der Pol for their valuable advice with respect to Chapter 3; Lorette Bosch-Padberg for the excellent way in which she designed the final version of the dissertation; Ellen Timminga, Hilde Tobi, Jos J. Adema and Carl P.M. Rijkes for their support and sympathetic ear in bleak moments; Dick van de Sar for the design of the user interface of the computer programme required for the research; Hanneke J.M. Lijklema for the excellent way in which she corrected the English version of the dissertation; the Centre for Biostatistics for providing me with time and space to complete my dissertation.

Finally I wish to thank my parents, whose support, love and confidence have made an indirect but important contribution to the realization of this dissertation.

Utrecht, December 1993

Paul Westers

CONTENTS

Chapter 1 Introduction and overview

1.1 Introduction 5

1.2 Test item bias 6

1.3 Item bias detection methods 8

  1.3.1 Unconitional methods 9
  1.3.2 Conditional methods 10
  1.3.3 Comparison of the item bias detection methods 12

1.4 Item bias detection based on item response theory and latent class analysis 13

  1.4.1 Item Response Theory 13
  1.4.2 Item bias and IRT models 14
  1.4.3 Latent class models 16

1.5 Research strategies for the explanation of item bias 17

1.6 Topic of the dissertation 18

1.7 Overview of the following chapters 20

Chapter 2 The examination of differential item functioning due to item difficulty and alternative attractiveness

2.1 Abstract 23

2.2 Introduction 23

2.3 A model for multiple-choice items that accounts for the selection of each alternative 25

2.4 The model written as an incomplete LCA model 27

2.5 Testing for DIF by related LCA models 29

  2.5.1 DIF in the latent response 30
  2.5.2 DIF in the attraction parameters 30

2.6 Parameters estimation and model testing 31

2.7 An empirical example 32

2.8 Discussion 35

Chapter 3 The estimation of the parameters of the solution-error response-error model with the use of subsets of items

3.1 Abstract 37

3.2 Introduction 37

3.3 The solution-error response-error model 40

  3.3.1 Constraints on model parameters 42

3.4 The division-by-item principle in the solution-error response-error model 43

3.5 The estimation of the parameters of the solution-error response-error model 46

  3.5.1 Traditional estimation method for the SERE model 47
  3.5.2 Restricted parameters 50
  3.5.3 Pseudo-likelihood theory 50
  3.5.4 The simultaneous estimation method 52
  3.5.5 Initial values 55
  3.5.6 Testing the SERE model 58

3.6 Applications of the DBI-principle 60

3.7 Discussion 63

Chapter 4 A simulation study of the solution-error response-error model

4.1 Introduction 67

4.2 The solution-error response-error model 67

4.3 Research questions of the simulation study 70

4.4 The simulated data 72

4.5 Results 74

  4.5.1 Small-sample behavior of pseudo-likelihood estimates 74
  4.5.2 Small-sample behavior of DIF detection 76
  4.5.3 The influence of the test length on DIF detection and parameter estimates 77
  4.5.4 The detection of DIF in the latent response and DIF in the attraction parameters 78
  4.5.5 Small-sample behavior of simultaneous detection of DIF in the latent response and DIF in the attraction parameters 79

4.6 Conclusion and discussion 82

Chapter 5 Generalizations of the solution-error response-error model

5.1 Abstract 85
5.2 Introduction 85
5.3 The generalized solution-error response-error model 87
5.4 Applications of the GSERE model 90
5.5 Identifiability 92
5.6 Estimation method 93
  5.6.1 Outer iteration 93
  5.6.2 Inner iteration 94
5.7 Goodness-of-fit test 95
5.8 An empirical example 96
5.9 Discussion 102

Chapter 6 Epilogue

6.1 Introduction 105
6.2 Summary 105
6.3 Future research 107

Appendices

A.1 The English version of Item 4 of the Second International Mathematics Study 109
A.2 The collapsed solution-error response-error model 109
A.3 LANPACO 110
A.4 Two items of the American Society of Clinical Pathologist (ASCP) Microbiology Test 111
A.5 The collapsed generalized solution-error response-error model 112

References 115

Nedertandse samenvatting (dutch summary) 127

Curriculum Vitae 131

Chapter 1

INTRODUCTION AND OVERVIEW

1.1 INTRODUCTION

One of the main issues in test theory and the practice of educational testing is differential item functioning (DIF) or item bias (Lord, 1980). Items in educational or psychological tests show DIF when the probability of a correct response among equally able test takers is different for various racial, ethnic, or gender subgroups (Mellenbergh, 1989). Although researchers have offered many methods for the detection of biased items, they have seldom offered explanations why items show DIF. Moreover, most of the methods are focused on the detection of biased items when models for binary (incorrect/correct) answers are used.

Recently, DIF detection research has addressed the differential functioning of all response alternatives in an item (e.g. Scheuneman, 1987; Schmitt, 1988). The main question in these studies is whether bias can be located in a differential selection of response alternatives in multiple-choice items. As an extension of the terminology of DIF, differential alternative functioning (DAF) exists when the attractiveness of the response alternatives of the item is different for equally able test takers. The attractiveness of an item represents the probability that a subject who does not know the correct answer will choose that alternative.

In this dissertation, the problem of both DIF and DAF detection in multiple-choice items is addressed. In Section 1.2 the issues of DIF and DAF will be discussed, whereas in Section 1.3 various item bias detection approaches are described. In this study the focus of attention is on one of these approaches: item bias detection based on item response theory. The approach will be discussed in Section 1.4. In Section 1.5 research strategies are described to explain why an item is biased. Section 1.6 deals with the statistical model on which the proposed item bias detection method will be based. Finally, in Section 1.7 a summary of the contents of the following chapters is given.

1.2 TEST ITEM BIAS

One of the most complex and confusing issues in educational and psychological assessments is test and item bias, where bias means that a subset of the total group of examinees (referred to as Focal group) responds in a different way to a specific test or item than the remainder of the population or another subgroup (referred to as Reference group) does. Generally, a test or item is biased if it functions differently for different groups. Bias can best be understood within the context of validity and fairness of tests for all persons. A test is biased or unfair if test scores, and hence the predictions or decisions based upon the scores of the test, are different for equally able subjects (Cleary, 1968). In other words, a test is unbiased and fair if the same predictions and decisions are made for equally able persons, regardless of group membership. Analogously an item is unbiased when subjects with the same ability have the same probability of responding correctly to the item. The following example will illustrate this.

In Kok (1982) bias in multiplication items was studied through a manipulation of the subjects' skill on possible bias factor. Of the administered multiplication items some items were written in the native language (i.e. Dutch), whereas for the other items Roman numerals were used. Moreover, the subjects were divided into two groups. Subjects from one group were trained with regard to Roman numerals, whereas the other group got no training. It is not surprising that the probability of responding correctly to a Roman numeral item for the trained group on the items is higher than that for the untrained group; this result would lead to the assumption that the item is biased. Since the difference in these probabilities may also be a result of a difference in ability, this assumption may be incorrect. If the test was to be used at an end-of-year examination of the subject's knowledge of multiplication techniques, the untrained group might receive low grades that did not reflect their real knowledge of multiplication techniques. The use of the test is then unfair and biased. On the other hand, if only the native language items were used for the examination and if there is no other biasing factor, the test would not be biased against the untrained group. Low scores would reflect the lack of knowledge of multiplication techniques.

The above example demonstrates two major points of bias. In the first place, fairness and bias are functions of the test used and they depend on the population measured and the use of measurement. Secondly, if a test is biased it does not mean that all the items of the test are biased. Generally, when a test is biased one or a few of the items are also biased. However, if there is no evidence that the test is biased, the items in the test can still be biased, because it might be the case that bias of one item is compensated by bias of another item.

As already observed, a single test item is biased if subjects with the same ability have a different probability of answering the item correctly. Since the probability of answering the item correctly is related to the difficulty of the item, DIF indicates that the difficulty of the item may be different for various equally able subgroups.

Generally, if significant differences in proportions of correct answers are found for equally able groups, DIF has been shown. It should be noted that if there is no evidence that the difficulty of an item is different for various equally able subjects, the item may still be functioning differently. Multiple-choice items, for instance, are functioning differently if the attractiveness of the incorrect alternatives (i.e. distractors) are different for various subgroups. Green, Crone and Folk (1989) stated that "although group differences in distractor choice have no effect on test scores, because all distractors are wrong, group differences might indicate that the item was functioning differently for the various subgroups. Items that have different meanings to different groups would seem to be biased in a very fundamental sense" (p. 148). Extending the terminology of DIF, an item shows differential alternative functioning (DAF) if the attractiveness of the alternatives of an item is different for equally able subjects.

Although DAF might have no consequences for the number of correct scores on the test, it can provide the test constructor with information about content areas that are problematic in terms of bias, which can then be accounted for in future test constructions. The next example of Veale and Foreman (1983) illustrates this.

A sample of 98 black and 412 white students from grade six were asked to pick the correct sentence from the following four sentences: (A) Janies takes her work seriously; (B) Janies work take too much time; (C) Working with books are my favourite thing; (D) Things people likes to do is their business. The p-values (i.e. proportion correct) were .45 for the black students and .73 for the white students. Further, in this item, ten percent of the white students were strongly attracted to distractor C, which was probably the most difficult distractor to eliminate, due to the juxtaposition of "books", "are" and "my". On the other hand, the black students were heavily drawn to distractor D; thirty percent of the blacks students have chosen that alternative. In the diagnosis of the source of the bias, distractor D seemed to be the source of the bias, because the construction "Things people likes to do ..." is one which occurs frequently in everyday "street" language of black children. Based upon this result, test-constructors might revise distractor D into "People does not like to work too hard" and change the stem of the item into "Pick out the sentence below that uses correct standard English". These revisions provide better directions to the student to discriminate between street language and correct standard English. They also focus the student on the exact purpose of the item. However, the intent of the original item has been maintained; the critical diagnosis of subject-verb agreement has been preserved.

Generally, the attractiveness of the alternatives depends on the insight a subject has into the solution of the item. Some wrong alternatives appeal to subjects who do not know the solution of the item; other distractors provide reasonable but wrong alternatives that might be chosen according to partial knowledge of the solution of the item. Since the insight a subject has into the solution of the item not only depends on the ability level of the subject, but also on the difficulty of the item, the attractiveness of the alternatives will also depend on the difficulty of the item. However, the probability of a subject responding correctly to an item not only depends on the difficulty of the item, but also on the attractiveness of the correct alternative. If the problem imposed by the item is relatively difficult, whereas at the same time the attractiveness of the correct alternative is great, a subject with a low ability still has a good chance of responding correctly to the item.

This means that a difference in functioning of a multiple-choice item may be caused by the item difficulty, the attractiveness of the alternatives, or both. Therefore, in the decision whether a multiple-choice item functions differently for equally able subjects, a distinction should be made between DIF and DAF. In this dissertation the issue of simultaneously detecting both types of item bias in multiple choice items will be addressed.

Before we discuss this issue in more detail (Section 1.6), we will first discuss methods for detecting biased items (Sections 1.3 and 1.4) and methods for explaining why an item is biased (Section 1.5).

1.3 ITEM BIAS DETECTION METHODS

At first sight the determination of DIF may seem to be simply a matter of comparing the conditional proportion of correct answers or the proportions of incorrect alternatives for the two groups. If significant differences in the proportion of correct answers are found for equally able subgroups, DIF has been shown to exist. Analogously, DAF has been shown to exist if significant differences in the conditional proportions of incorrect alternatives are found for equally able groups.

Many detection methods have been proposed to find biased items (Baron, 1988; Berk, 1982; Osterlind, 1983; Rudner, Getson & Knight, 1980a, 1980b). Generally, these methods can be divided into a group of unconditional methods and a group of the earlier conditional methods. The difference between the two groups can be described as follows (Mellenbergh, 1985). In both detection methods bias is regarded as an interaction between groupmembership and item difficulty: The differences in item difficulty between groups is not constant for all items, and items that deviate from the general trend are considered to be biased. Conditional detection methods consider an item to be biased if the item is functioning differently for subjects with the same ability. In the unconditional detection methods, however, the condition of equal ability of the subject is not considered and are therefore based on a incorrect definition of DIF.

In order to give a complete (historical) survey of the item bias detection methods, both unconditional detection methods and conditional detection methods will be briefly described and discussed below. For a complete and detailed discussion of these methods, the above mentioned references and the papers of Shepard, Camilli and Averill (1981) and Shepard, Camilli and Williams (1984, 1985) are recommended.

1.3.1 Unconditional methods

There are two groups of unconditional bias detection methods: methods based on analysis of variances (ANOVA) and those based on transformed item difficulties (TID).

In the ANOVA approach it is incorrectly assumed that the interaction effect of groups by items on the variation in item scores, is a valid indicator of DIF. If the null hypothesis of no significant interaction effect of groups by items is rejected, The existence of DIF is inferred. Since for groups of unequal ability, however, item by group interaction will occur in completely unbiased tests, it cannot be concluded that the item shows DIF merely because the null hypothesis is rejected. Hunter (1975) illustrated this problem by showing that items of different overall difficulty will always "...show an item by group interaction in any situation in which the two groups differ in achievement level" (p. 10). Examples of DIF detection studies with the ANOVA methodology are those of Cardall and Coffman (1964) or Cleary and Hilton (1968) .

In the delta plot method of Angoff (1972; Angoff & Ford, 1973), also called the transformed item difficulty (TID) approach, two sets of item p-values are computed, one for the Focal group and one for the Reference group. Each p-value is transformed to normal deviates. Then for each item in the test the pair of normal deviates is plotted in a bivariate scattergram and a baseline is drawn from the lower left to the higher right. This baseline represents the best fit of the scattergram and it will be used to gauge the amount of bias. The distance between a particular item and the baseline will then be used to indicate the magnitude of DIF. Items which, according to a method of outlier or residual analysis, deviate greatly from this line are regarded as showing DIF. Sometimes the delta plot method is used as a post-hoc procedure to ANOVA.

One disadvantage of the delta plot method is that it will produce spurious evidence of DIF unless the items are all equal in discrimination or the groups being compared do not differ in average performance (Angoff, 1982; Hunter, 1975; Lord, 1977). To correct for these sources of error, Angoff (1982) proposed to adjust the delta plot method for item-test correlation or to match the groups on ability beforehand. In the latter case, the adjusted delta plot method would be a conditional item bias detection method.

In the papers of Echternacht (1974), Jensen (1980), Rudner, Getson and Knight (1980a, 1980b) and Stricker (1981), several variations of the delta plot method can be found.

1.3.2 Conditional methods

Both the delta plot method and ANOVA do not consider the ability level of the subjects directly, which may lead to questionable decisions about the presence of DIF in items. For instance, interactions of groups by items on the variations in item scores can occur in any test regardless of DIF (Hunter, 1975). When a test in English is administered to students from different school grades, the use of ANOVA would lead to the decision that the items show DIF. However, this may be not correct, because school grade may be associated with the response to the test. In this case a latent trait, such as an English language ability, may be used to explain the significant interaction effect of groups by items.

In the remainder of this chapter, bias detection methods will be discussed that account for the different ability of the subjects. In these methods the total number of correct scores on a test is used as a measure of ability, and an item is defined to be functioning differently, if for all subjects at the same level of total number correct scores, the proportion of correct responses or the proportion of incorrect alternatives is different for various groups of subjects.

These conditional bias detection methods can be divided into four groups: methods based on chi-square statistics, on factor analysis, on distractor analysis and on item characteristics curves (ICC).

In Chi-square methods (Marascuillo & Slaughter, 1981; Mellenbergh, 1982; Scheuneman, 1979) the subjects are divided into a number of score levels according to their scores on the test under study. To examine whether an item shows DIF, the proportions of correct responses for the Focal and the Reference group are then compared within each score level. In this technique for each score level an expected set of proportions of correct responses is calculated for the two-way group membership by response contingency tables, assuming independence. Then for each table Pearson chi-square statistics are calculated and summed up across score levels. If the calculated chi-square is statistically significant, the item shows DIF. Otherwise, the item does not show DIF. Essential in this technique is the absence of the use of the distribution of correct responses across ability levels. Since there is a wide variety available of statistical DIF detection methods based upon chi-square procedures, only the one that has received a great deal of attention lately will be discussed: the Mantel-Haenszel (MH) statistic.

The Mantel and Haenszel (1959) statistic, adopted for DIF analysis by Holland and Thayer (1988), also compares the item performance of a Reference and Focal group, across different score-levels. The MH procedure is as follows. First the subjects in the Reference and Focal group are divided into subgroups at different score-levels. Then for each score-level for both groups the odds ratio of the p-value is computed. Finally, it is tested whether for some of the score levels, the odds ratio for the Focal and for the Reference group differs by a constant factor α

$$q_k/(1-q_k) = \alpha\, p_k/(1-p_k), \qquad \text{for all } k = 1,\ldots,K$$

in which K denotes the number of score levels and p_k and q_k denote the p-values in the Focal and in the Reference group for each score level k. When α is not significantly different from the value one, it may be concluded that the item shows no DIF, which means that both groups perform equally well on the item when their abilities are taken into account.

Let A_k and C_k be defined as the number of subjects in the Reference group and Focal group at score-level k who answered the item correctly. Finally, let B_k and D_k be similarly defined for the number of subjects who answered the item incorrectly. The MH odds ratio estimator is then defined as

$$\alpha_{mh} = \Sigma(A_kD_k/N_k) \,/\, \Sigma(B_kC_k/N_k)$$

where Σ is the summation over all score-levels k ($k=1,\ldots,K$) and N_k is the total number of subjects at score level k. To test whether the observed α_{mh} is significantly different from one, Mantel and Haenszel (1959) proposed the chi-square statistic MH-CHISQ with one degree of freedom,

$$\text{MH-CHISQ} = (|\, \Sigma A_k - \Sigma E(A_k)| - 0.5)^2 \,/\, \Sigma \operatorname{var}(A_k),$$

where $E(A_k)$ and $\operatorname{var}(A_k)$ denote the expected value and variance of A_k under the null hypothesis that the factor α_{mh} equals one. According to Holland and Thayer (1988) this chi-square test offers a powerful test of the null hypothesis.

To summarize, the use of the MH statistic for DIF analysis can be viewed as an extension of the ideas behind the chi-square procedures of Marascuillo and Slaughter (1981), Mellenbergh (1982), and Scheuneman (1979). It provides not only a more powerful test for examining DIF, but it also produces a measure of the degree of DIF showed by the studied item. Additional details concerning the MH technique as used in educational testing can be found in Holland and Thayer (1988), Linacre and Wright (1986), and Raju, Bode and Larsen (1987).

In Bias detection strategies based on factor analysis an attempt is made to explain the test performance by underlying factors (i.e. dimensions or traits). Different sets of factor loadings for a Focal and a Reference group may indicate that the two groups are not responding to the items in the same manner. If that is the case, a test would be considered biased and the item with the largest difference in factor loadings shows the most pronounced DIF. Examples of bias detection methods based on factor analysis can be found in Green (1976), Green and Draper (1972) and Merz (1973, 1976). The approaches of Green and Draper are attractive in the sense that item variances are partitioned into culture-specific and culture-common sources, whereas Merz's approach has the advantage that variances attributable to racial, ethnic or gender factors can be distinguished.

Despite the fact that the factor analytical techniques deal with the underlying abilities of the examinees, these techniques are not recommended for the analysis of test bias or DIF. The main reason is, that the decision problems that beset factor analysis in general are increased when these techniques are applied to item bias. See Rudner and Convey (1978) and Rudner, Getson and Knight (1980a) for a discussion of these problems.

The distractor response analysis approach focuses on the differential attractiveness of the distractors. If a significant test reveals that the equally able Focal and Reference group are differently attracted to an item distractor, the null hypothesis of no difference in the groups' relative frequency distribution for distractors may be rejected, and it may be concluded that DAF exists. Just like the chi-square approach, in the distractor strategy a conditioned difference in proportions is examined. In the papers of Green, Crone and Folk (1989) and Veale and Forman (1983) some applications of this approach are given. For instance, Green, Crone and Folk perform a loglinear analysis of the subgroup by number correct score by distractor contingency table for each item. They test the null hypothesis that the interaction between subgroup and distractor is not needed to explain the observed item responses. If the data cannot be explained without the interaction of subgroup by distractor, they define the item to show DAF. In a similar approach, Veale and Foreman define an unconditional model that incorporates parameters representing (a) achievement differences across groups and (b) differences in alternative difficulty. Their method also provides information about the source of the bias, so the item may be revised to eliminate the bias, rather than eliminating it.

Finally, the ICC approach for the detection of DIF is derived from the item response theory (IRT). Since the contents of this dissertation is focused on this approach, it will be discussed separately in Section 1.4.

1.3.3 Comparison of the item bias detection methods

Comparative studies of the above item bias detection methods (Ironson & Subkovak, 1979; Rudner, Getson & Knight, 1980a; Shepard, Camilli & Averill, 1981; Shepard, Camilli & Williams, 1984, 1985) show that the examination of item bias is improved when methods take the ability into account. Furthermore, methods based on IRT appear best, the delta plot method is poorest and the chi-square methods are in between (Baron, 1988). The delta plot method is acceptable, however, when the difference in ability between the Focal group and the Reference group is small and DIF is more the result of the difficulty of the item rather than of its discriminating power. Furthermore, on both logical and empirical grounds (Ironson & Craig, 1982; Shepard, Camilli & Averill, 1981), Scheuneman's (1979) earlier chi-square method has to be replaced by the full chi-square procedure. Finally, the IRT method is preferred theoretically (Lord, 1977; Petersen, 1977), but in some cases it is not better than the chi-square method. The Monte Carlo study of Rudner, Getson and Knight (1980b), for example, shows that both the three-parameter logistic model and the chi-square method with five score-levels produced fairly accurate results under all investigated conditions. Moreover, comparative studies of the Mantel-Haenszel procedure and a procedure based on IRT models showed that, under the Rasch model, the identity of ICCs across groups of subjects implies that the MH null hypothesis is met (Holland & Thayer, 1988). However, identity of the ICCs does not imply that the MH null hypothesis is met when the item response functions are monotonic and where local independence holds (Zwick, 1990). For example, when each item has the same item response function in the Focal and the Reference group and the ability distributions are ordered, the MH procedure will show DIF which, moreover, favors the group with higher ability.

1.4 ITEM BIAS DETECTION BASED ON ITEM RESPONSE THEORY AND LATENT CLASS ANALYSIS

As mentioned before, in this section DIF detection will be approached from an item response theory (IRT) perspective. In Section 1.4.1 the basic ideas underlying IRT will be discussed. Then, in Section 1.4.2, DIF detection methods based on IRT models will be discussed. Finally, the latent class model will be discussed.

1.4.1 Item Response Theory

Over the last twenty years DIF detections methods have been proposed that are based on IRT. An IRT model describes the probability of a correct response to a dichotomous item as a mathematical function of person and item characteristics. These functions are known as the item characteristic curves (ICCs). The simplest IRT model is the following Rasch (1960/1980) model, which has only one item parameter (the difficulty parameter):

$$P(Y_j=1|\theta) = \exp(\theta-\delta_j)/(1+\exp(\theta-\delta_j)) \, , \tag{1.1}$$

in which $P(Y_j=1|\theta)$ is the probability of a correct answer to item j of a subject with ability θ and δ_j is the difficulty parameter of item j. The Rasch model is a special case of the following three-parameter logistic model (Birnbaum, 1968)

$$(1.2) \qquad P(Y_j=1|\theta) = \gamma_j + (1-\gamma_j)\exp(\alpha_j(\theta-\delta_j))/[1+\exp(\alpha_j(\theta-\delta_j))] ,$$

where α_j and γ_j are the item discrimination and guessing parameter, respectively. The guessing parameter γ_j can be seen as the probability of a correct response to item j if $\theta \rightarrow -\infty$. Examples of other IRT models are the two-parameter logistic model (Maxwell, 1959), the partial credit model (Masters, 1982), the linear-logistic model (Fischer, 1973; Scheiblechner, 1972) and the normal ogive model (Lord, 1980; Lord & Novick, 1968). This list is not exhaustive. The only IRT model addressed in this dissertation is the Rasch model.

One of the common assumptions in IRT is the unidimensionality of the ability space. Unidimensionality is understood to mean that the number of abilities that accounts for the subjects' performance on the test equals one.

Another common assumption in IRT is the local independence of items, which means that for a fixed subject, response on any item is independent of the response on the other items in the test. Consider, for example, a test consisting of k items, where each has two response categories: incorrect (0) and correct (1). Further, let Y_j denote the observed response on item j ($j=1,\ldots,k$) and let the response pattern of a subject on all k items be denoted by $\mathbf{Y}=(Y_1,\ldots,Y_k)$. Then the probability $P(\mathbf{Y}=\mathbf{y}|\theta)$ that a subject with ability θ will have response pattern $\mathbf{y}$ on the k items is equal to

$$(1.3) \qquad P(\mathbf{Y}=\mathbf{y}|\theta) = P(Y_1=y_1|\theta) \ldots P(Y_k=y_k|\theta) .$$

Thus local independence means that the simultaneous probability is the product of the independent item probabilities.

One of the main features of the Rasch model is that the equations for the estimation of the item parameters can be obtained independently of the ability parameters, and vice versa (Fischer, 1974, 1987; Glas, 1989). Finally, Fisher's information can be used as a measure for the accuracy of a test at ability level θ, because it is inversely proportional to the asymptotic standard error of the maximum likelihood ability estimates at this ability level (Lord, 1980).

1.4.2 DIF and IRT Models

With IRT an unbiased item can be defined as follows: An item is unbiased if the ICCs for the various groups are identical. Conceptually, this definition and the definition of DIF as mentioned in Section 1.2 are alike. Mathematically, this definition implies that an item shows DIF if the ICC is different for any subgroup. This means that, in the Rasch model, the probability of a correct response to item j in subgroup i (i=1,...,g) is equal to

$$P(Y_{ij}=1|\theta) = \exp(\theta-\delta_{ij})/[1+\exp(\theta-\delta_{ij})] \,, \tag{1.4}$$

in which $P(Y_{ij}=1|\theta)$ is the probability of a correct answer to item j of a subject with ability θ from subgroup i, and δ_{ij} is the difficulty parameter of item j in subgroup i. If the item is unbiased then the difficulty parameter δ_{ij} is equal over all subgroups, i.e. $\delta_{ij} = \delta_j$ for all subgroups i.

Since for a chosen IRT model the ICCs are determined by the item parameters, the statistical question in DIF detection is whether the item parameters for the Reference and Focal groups differ significantly. An item is said to be biased if the difference between the item parameter estimates for the Reference and Focal groups is significant under a certain IRT model. In view of the variety of IRT models and approaches to parameter estimation and hypothesis testing, there are different approaches for the detection of significant differences between ICCs. Some of these methods are based on marginal maximum likelihood estimation (Thissen, Steinberg & Gerhard, 1986; Thissen, Steinberg & Wainer, 1993), others on the loglinear formulation of the Rasch model (Kelderman, 1989; Kelderman & Macready, 1990) or on generalized least square estimation (Muthén & Lehmann, 1985). In all three approaches likelihood ratio tests are used for the evaluation of the significance of the observed differences between two groups. A fourth detection method is also based on marginal maximum likelihood estimation, but here the ratios of the parameter estimates to their standard errors are used to test whether the item parameters differ significantly across groups (Bock, Muraki & Pfeiffenberger, 1988; Muraki & Engelhard, 1989). In Thissen, Steinberg, and Wainer (1993) a good survey of these detection methods is given. They conclude that each of the four methods implements estimation and hypothesis testing for distinct subsets of IRT models and that they perform as expected when the model is appropriate for the data. Therefore, the choice between these four methods must be made with respect to the assumptions of the model.

In the literature there are also entirely descriptive IRT based bias detection methods (Hambleton & Swaminathan, 1985; Linn, Levine, Hastings & Wardrop, 1981). In most of these methods the area between the ICCs of the Focal and Reference group is used to examine the existence of DIF (Raju, 1988; Rudner, Geston & Knight, 1980a, 1980b; Shepard, Camilli & Averill, 1981; Shepard, Camilli & Williams, 1984). The paper by Raju offers formulas for computing the exact area between the two ICCs of the one-, two-, and three-parameter models under the restriction that the guessing parameters are equal for both ICCs. In the other references the area between the ICCs of the two groups is computed by integration over an ability interval. The use of the area between two ICCs for DIF detection may be worthwhile. However, most of these methods are not associated with any inferential statistics that can be used for DIF detection.

1.4.3 Latent Class Models

In most cases it is assumed that the ability parameter θ is a *continuous* latent variable. However, there are also response models for which it is assumed that the latent space is *discrete*. These models are known as the latent class analysis (LCA) models. Thissen and Mooney (1989) wrote: "IRT models are developed as tools for measurement, whereas LCA models are presented primarily as structural models for observed item response data. ... The crucial difference between IRT models and LCA models is that IRT models are based on the relationship of the probability of a particular item response with a continuous latent variable" (p. 300-301).

Each point in the discrete latent space corresponds to a latent class. The probability of a positive response of a subject to each of the items is completely specified by these latent classes and the conditional probabilities of a correct response to the item, given the subject's membership in the latent space. In contrast to IRT models with continuous latent traits, no specific functional form is assumed for the conditional probabilities.

Just as in the IRT case, local independence is assumed in LCA, which means that within a latent class the responses to items are all independent. For example, consider again a test consisting of k items, where each has two categories: incorrect (0) and correct (1), respectively. Let Y_j denote the observed response on item j (j=1,...,k) and let the response pattern of a subject on all the k items be denoted by the vector $\mathbf{Y}=(Y_1,...,Y_k)$. Further, let T denote the number of latent classes, X be the random variable associated with the latent classes, $P(X=x)$ be the probability that a subject will be in the x*th* latent class (x=1,...,T) and $P(Y_j=y_j|X=x)$ be the conditional probability that a subject will score response category y_j (y_j=0,1) given that the subject is in the x*th* latent class. Then the probability $P(\mathbf{Y}=\mathbf{y})$ that a subject will have response pattern $\mathbf{y}$ on the k items is in a LCA model equal to

$$(1.5) \qquad P(\mathbf{Y}=\mathbf{y}) = \sum_x P(X=x)\, P(Y_1=y_1|X=x) \ldots P(Y_k=y_k|X=x) .$$

in which Σ_x is the summation over all latent classes x (x=1,...,T).

For the detection of bias, Model (1.5) has to be extended to include a variable which denotes group membership. Let i (i=1,...,g) denote the group membership of a subject, then the probability that a subject from group i will have response pattern $\mathbf{y}$ is equal to

$$P(Y=y|i) = \sum_{x} P(X=x|i)\, P(Y_1=y_1|i,X=x) \ldots P(Y_k=y_k|i,X=x)\,, \tag{1.6}$$

in which $P(X=x|i)$ is the probability that a subject from group i will be in the x*th* latent class and $P(Y_1=y_1|i,X=x)$ is the conditional probability that a subject will score in response category y_j, given that the subject is in the x*th* latent class and in group i.

If the marginal distribution of subjects over the latent classes is equal over all groups, then $P(i,x) = P(x)$ for all groups i. If the association between item j and the latent classes is equal in all groups, then $P(Y_j=y_j|i,X=x) = P(Y_j=y_j|X=x)$ for all groups i and latent classes x. if this is the case, then item j shows no DIF (Clogg & Goodman, 1985; Mellenbergh, 1985, 1989). However, item j shows DIF if the marginal probability $P(Y_j=1,i)$ is not equal in all groups i.

1.5 RESEARCH STRATEGIES FOR THE EXPLANATION OF ITEM BIAS

Traditionally, most of the previous mentioned item bias detection methods focus on the detection of biased items. Only a few item bias detection methods also try to examine why an item is biased. For example, item bias detection methods based on distractor analysis not only yield information about whether an item is biased, but it also provides information about which alternative of the item was likely to be responsible for the bias (Green, Crone & Folk, 1989; Veale & Foreman, 1983). However, these item bias detection methods are not among the best detection methods. The method of Green, Crone and Folk, for instance, is not based on an IRT model and the method of Veale and Foreman does not control for ability.

Of course there are several research strategies to explain item bias, but not all of them have been applied in empirical research. For instance, Ackerman (1992) suggested to explain item bias from a multidimensional perspective. The idea behind his strategy is that "if two different groups of subjects have different underlying multidimensional ability distributions and the test items are capable of discriminating among the levels of abilities on these multiple dimensions, then any unidimensional scoring scheme has the potential to produce item bias" (p. 67). Ackerman's strategy can be viewed as an extension of the strategies of Kok (1988) and Shealy and Stout (1991). Kok presented a mathematical model to explain how DIF can occur because of multidimensionality, and Shealy and Stout presented a detailed theoretical analysis of DIF from a similar perspective as Kok.

In Smith and Camilli (1988) the question "What caused the bias?" is replaced by the question "Who caused the bias?". The idea behind their strategy is that it is not always the group as a whole that causes the bias, but a recognizable subgroup within the disadvantaged group. For example, suppose that the correct answer of an item is B, but that a highly plausible distractor is

C. Impulsive subjects may then be much more likely to select C and will not look further. If impulsiveness is more prevalent in one group than in the other, the item may show DIF. So, if DIF can be explained by certain characteristics of some subjects in the disadvantaged group, then it should be possible to identify these aberrant subjects and examine how these subjects differ from the other subjects in the group. Explaining DIF from the perspective of the subject is not usually done. However, it is a strategy which should not be neglected, because DIF involves not only the test items, but also the test takers (Linn & Harnisch, 1981).

In order to summarize, bias may be explained by either qualitative, correlational, quasi-experimental or experimental strategies (Mellenbergh & Kok, in press). The differences between these strategies can be described as follows. In the qualitative strategy a study is made of either the item content or the subjects' cognitive process when answering the item, whereas in the correlational strategy the relations between the item responses and variables of interest are studied. Examples of these variables may be subjects characteristics or item characteristics. Furthermore, in quasi-experimental studies either the predetermined groups of subjects or the predetermined groups of items are compared. Finally, in experimental studies the subjects characteristics or the characteristics of the items are manipulated. Examples of the applications of these strategies can be found in Lucassen and Evers (1984), Scheuneman (1987), Subkoviak, Mack, Ironson, Craig (1984) and Van der Flier (1982) respectively.

A problem of the above item bias research strategies is that all of them are follow-up analyses. One of the item bias detection methods has to be used to detect the biased items, and only then one of the research strategies can be used to examine "What" or "Who" caused the bias or to examine the bias from a multidimensional perspective. This makes the examination of item bias not only difficult and problematic, but also inefficient. Therefore, a development of DIF detection methods that give more information about the nature of DIF may be appreciated.

1.6 TOPIC OF THE DISSERTATION

In the previous sections several questions were raised. In the first place, bias in multiple-choice items may be caused by a combination of the difficulty of the item and the attractiveness of the alternatives. In nearly all of the existing item bias detection methods, however, only one type of bias at the time is considered.

Secondly, although item bias can be defined and biased items can be detected, only minor attention has been given to the explanation of the bias factor. In practice, biased items were removed from the test and the test was claimed to be fair with respect to the groups investigated. The advantages of knowing why an item is biased, for test construction, has been neglected for a long time. However, in the last decade researchers have become aware of the fact that knowledge about factors causing bias can prevent the occurrence of biased items in new tests. This consideration has resulted in several research strategies for the examination of item bias. These research strategies, however, are not only difficult and problematic to apply, they also consider one type of item bias at the time and are therefore inefficient. In this study we will propose an item bias detection method that makes it possible to test whether an item shows DIF, DAF, or both.

As far as we know there are only a few IRT models that make it possible to distinguish between DIF caused by the difficulty of the item and DIF caused by the attractiveness of the alternatives. A model that may come first to mind is the three-parameter logistic model. However, the guessing parameter (i.e. γ) of this model denotes the attractiveness of the correct alternative, whereas ($1-\gamma$) denotes an overall attractiveness of the distractors. So the three-parameter logistic model does not account for differences between the attractiveness of the different distractors. Other models which can be used are the multiple-choice model of Thissen and Steinberg (1984), the model of Lord (1983) or the solution-error response-error model of Kelderman (1988). All three models not only concentrate on the observed responses, but also on the process leading to these observed responses. In this way it should not only be possible to test whether an item is biased, but also to get more information why an item is biased. For example, before a subject responds to a multiple-choice item, (s)he must first recognize and solve the problem imposed by the item, and then choose one of the alternatives. At each level of this process there may be danger of bias. For instance, the probability that a subject can recognize and solve the problem depends on the difficulty of the item, which may vary across different subgroups. If this difficulty is not equal for different subgroups, then the item shows DIF. Furthermore, whether or not the subject has solved the problem, (s)he has to select one of the alternatives, and this may depend on the attractiveness of the alternatives. If the attractiveness of the alternatives differs for subjects from different subgroups, then the item shows DAF. In this study the solution-error response-error model (SERE) of Kelderman (1988) is used to examine both types of bias. Since the SERE model will be formulated and discussed more extensively in Chapter 2, only a brief description of the model is given here.

The SERE model is a loglinear Rasch model with latent classes and can be regarded as a two-process model. The first process determines whether or not a subject will be able to solve the problem imposed by the item. For this process, in the SERE model a distinction is made between a "Know" state, in which the subject has complete knowledge of the solution to the item, and a "Don't know" state. The probability that the subject is in the "Know" state rather than the "Don't know" state is assumed to be governed by the Rasch model. The responses of the solving process will be referred to as idealized responses or latent responses.

In the second process the answer of the subject on the item is determined. Whether or not the subject is in the "Know" state, (s)he has to choose one of the alternatives of the item. If the subject does not know the solution to the item, the choice of an alternative may depend on the attractiveness of the alternatives, which may be different for different alternatives, including the correct one. On the other hand, if the subjects is in the "Know" state, it may be expected that the correct answer will be chosen. However, the subject may choose also one of the distractors because of a writing error.

The observed responses are the result of the second process. In the SERE model the relationship between the latent responses and the observed responses are modelled by conditional probabilities. Since these conditional probabilities indicate the attractiveness of the alternatives, we will refer to these conditional probabilities as attraction parameters.

The SERE model can be seen as Macready and Dayton's (1980) extension of Goodman's (1975) model for scaling response patterns, but in which the deterministic Guttman (1950) model is replaced by the Rasch model. The SERE model is akin to the latent trait models of Lord (1983) and Thissen and Steinberg (1984). In these models it is assumed that subjects in the "Don't know" (Thissen & Steinberg, 1984) or "Undecided" (Lord, 1983) latent state arrive at an observed response by guessing. However, the SERE model is a more general model (Kelderman, 1988).

Furthermore, the SERE model represents one of the efforts to relate IRT models to LCA models. Other efforts were made by Bock and Aitkin (1981), Dayton and Macready (1980), Formann (1985), Kelderman and Macready (1990), Mislevy and Verhelst (1990), Rost (1990, 1991) and Yamamoto (1987, 1988). In all these models an attempt is made to combine the advantages of the IRT and LCA models into one single model so that more information on the knowledge of the subject can be obtained. For example, with LCA models it is possible to make a statement about the subjects' cognitive structure of understanding and misunderstanding a certain ability, such as arithmetic, foreign language and so on. Furthermore, LCA models have the advantage that the theory for maximum likelihood estimation and likelihood-ratio testing are well developed. On the other hand, with IRT models it is possible to attach scale values to subjects that represent the ability of the subject. The combination of the advantages of the IRT models and the LCA models into one model is very interesting for the study of DIF.

1.7 OVERVIEW OF THE FOLLOWING CHAPTERS

In Chapter 2 we will show that the SERE model can be used for examining DIF in multiple-choice items through a combination of the usual notion of DIF for correct/incorrect responses and information about DIF contained in each of the alternatives. In the method proposed incomplete latent class models are used to examine whether DIF is caused by the attractiveness of the alternatives, the difficulty of the item, or both.

As Kelderman (1988) has shown, parameter estimates can be computed when the programs LCAG (Hagenaars & Luijkx, 1990) and LOGIMO (Kelderman & Steen, 1988) are used. The underlying estimation method of these programs, however, becomes unpractical when the number of items is large. In Chapter 3 a method of parameter estimation is described that is based on dividing the whole item set in several subsets. We will show that, dependent on the number of items in each subset, the parameters of the SERE model can be estimated much more efficiently, both in terms of computer storage and processing time needed. Since information about the joint relationships among the items may be lost when the set of items is divided into subsets, the estimators of the parameters will, however, not be efficient.

Chapter 4 contains a simulation study of a DIF detection method based on the SERE model with an examination of the estimation method introduced in Chapter 3. The main questions considered are: (1) Can DIF still be detected if the number of items or the number of subjects is small?; (2) How do the values of the parameter estimators differ from the true model?; (3) Is this deviation consistent in the sense that the differences tend to decrease when the number of subjects increases?

The SERE model as described in Kelderman (1988) and Chapter 2 deals with a one-dimensional continuous latent trait. The production of one alternative response may, however, require quite another ability from the subject than the production of another answer. Besides, some responses may require the repeated application of an ability, whereas others may require only a single application of that same ability. This would mean that, although a multiple-choice item has one correct alternative, incorrect responses might often be chosen after cognitive activities similar to those necessary to arrive at the correct response. Therefore, in Chapter 5 the SERE model is generalized to a multidimensional polytomously scored latent response model. When this generalized SERE model is used, it is not only possible to detect both types of DIF, but also to explain DIF according to the ideas of Ackerman (1992), Kok (1988) and Shealy and Stout (1991). However, this point will not be further pursued in this dissertation.

Chapter 6 contains the summary of this dissertation as well as an overview of features of the on the SERE model based DIF detection methods that need further investigation.

Chapter 2

THE EXAMINATION OF DIFFERENTIAL ITEM FUNCTIONING DUE TO ITEM DIFFICULTY AND ALTERNATIVE ATTRACTIVENESS*

2.1 ABSTRACT

A method for analyzing test item responses is proposed to examine differential item functioning (DIF) in multiple-choice items through a combination of the usual notion of DIF for correct/incorrect responses and information about DIF contained in each of the alternatives. In the method proposed incomplete latent class models are used to examine whether DIF is caused by the attractiveness of the alternatives, difficulty of the item, or both. DIF with respect to either known or unknown subgroups can be tested by a likelihood ratio test statistic which is asymptotically distributed as a chi-square random variable.

2.2 INTRODUCTION

Items in educational or psychological tests may show differential item functioning (DIF). This means that the probability of a correct response among equally able test takers is different for various racial, ethnic, or gender subgroups. An educational or psychological test consisting of many items with significant DIF may be unfair for certain subgroups, and it is important to identify these items, so that they can be improved or deleted from the test. Many DIF detection methods have been proposed since Binet and Simon (1916) drew attention to this problem. Reviews of previous DIF (also called item bias) detection methods are given by Berk (1982), Osterlind (1983), and Rudner, Getson and Knight (1980a).

---

* This chapter is a slightly revised version of P. Westers & H. Kelderman (1992). Examining Differential Item Functioning Due to Item Difficulty and Alternative Attractiveness. ***Psychometrika*, *57***, 107-118. Printed by permission of the publisher.

In the last decade, the DIF detection methods have been improved to provide a better basis for matching on ability. In various methods the number correct score of the test has been used for this ability matching (Holland & Thayer, 1988; Mellenbergh, 1982; Scheuneman, 1979). Recently, DIF detection methods have been proposed which are based on item response theory (IRT) (Baker, 1977; Lord, 1980; Mellenbergh, 1989; Muthén & Lehmann, 1985; Wright, Mead, & Draba, 1975). Thissen, Steinberg, and Wainer (1993) give an overview of IRT-based DIF detection methods and demonstrate their use. They also discuss DIF detection methods which can be used with multiple choice items, where the response alternatives are also potential sources of DIF.

Green, Crone, and Folk (1989) focus on the differential attractiveness of the incorrect responses (or "distractors"). If a particular distractor is more attractive to subjects from one subgroup than from another, Green et al. conjecture that "...the item probably means something different to the different groups" (p. 147). They perform a loglinear analysis of the subgroup x score group x incorrect response contingency table for each item, to detect distractors that are more popular in one subgroup than in another. A similar approach of Veale and Foreman (1983) is based on the notion that examinees' responses to the incorrect alternatives provide more and better information concerning DIF than their responses to the correct alternative. Their model, called the overpull probability model, incorporates parameters representing (a) achievement differences across groups and (b) differences in alternative difficulty. Their proposed method also indicates the likely source of the bias so that the item may be revised to eliminate the bias rather than discarding the item. The methods proposed by Green et al. and Veale and Foreman have certain drawbacks; the Green et al. method, for example, is not based on an IRT model and the Veale and Foreman method does not control for ability. In the DIF detection method proposed in this chapter these two problems are avoided.

Another source of DIF in multiple choice items deals with the differential difficulty of the problem to be solved. An item may show DIF if it is more difficult for some subgroup than for others, even if they are equally able on the trait of interest (Lord, 1980; Rudner, Getson, & Knight, 1980a). In this chapter a DIF detection method is described that separates both sources of bias. In the proposed method, a distinction is made between a "Know" state in which the subject has complete knowledge of the answer and a "Don't know" state. Furthermore, it is assumed that if the subject is in the "Know" state, (s)he will give the correct answer. Here the probability that the subject is in the "Know" state is assumed to be governed by the Rasch (1960/1980) model. If the subject is in the "Don't Know" state, (s)he will choose the most attractive alternative, where the attractiveness of an alternative may be different for various alternatives, including the correct one.

The proposed DIF detection method differs from that of Thissen, Steinberg, and Fitzpatrick (1989), who distinguish between a "Don't know" state and a state in which the subject has partial or complete knowledge of the answer. In the "Don't know" state, the subject guesses the answer as before, but in the "Partial knowledge" state (s)he may select a response alternative according to response probabilities that are governed by Bock's (1972) nominal response model.

The proposed method is simpler than that of Thissen, Steinberg, and Fitzpatrick (1989). This simplicity has two advantages. In the first place, the method proposed here contains fewer parameters; for example, for a four-choice item the proposed model has five item parameters, while the model of Thissen et al. has fourteen. Obviously, if the sample is not very large, the parameters of the model by Thissen et al. cannot be estimated reliably. So, in that case, one may be inclined to "buy information by assumption" and use the simpler model. Secondly, the proposed model can be easily formulated as a latent class analysis (LCA) model (Kelderman, 1988). LCA models have been used extensively for measurements in sociology, psychology, and education (Clogg, 1981). There is a well-developed theory for maximum-likelihood estimation and likelihood-ratio testing of the LCA models (Goodman, 1978; Haberman, 1979; Lazarsfeld & Henry, 1968). By comparing the fit of different LCA models, DIF in the attraction of the alternatives and DIF in the parameters of the Rasch model can be tested separately (Kelderman, 1989; Kelderman & Macready, 1990). The model can also be extended to latent classes, so that the subgroups for which an item shows DIF may be latent.

A model for multiple choice items is developed below and formulated within the latent class framework. Different models for the detection of DIF are formulated, including a provision for the definition of the subgroup as a latent variable. A computationally efficient estimation method is described and illustrated with empirical data.

2.3 A MODEL FOR MULTIPLE-CHOICE ITEMS THAT ACCOUNTS FOR THE SELECTION OF EACH ALTERNATIVE

Let us suppose that each subject, randomly drawn from a population of subjects, responds to k test items, where the answer to item j may be any one of the r_j responses, denoted by y_j ($y_j=1,...,r_j$). Let x_j indicate the latent response of the subject, taking values $x_j=1$ if the subject is in the "Know" state (i.e. the subject has complete knowledge of the answer), or $x_j=0$ if the subject is in the "Don't know" state. The random variables associated with y_j and x_j are denoted by Y_j and X_j ($j=1,...,k$), respectively.

The relationship between the latent response x_j and the observed response y_j is described by the conditional probability

$$\Phi^{X_jY_j}_{x_jy_j} = P(y_j|x_j)\,, \tag{2.1}$$

in which the superscripts, in symbolic notation, indicate that the random variables X_j and Y_j are involved in the conditional probability. For the sake of simplicity, the notations y_j, x_j, et cetera in the probabilities are used for $Y_j=y_j$, $X_j=x_j$, et cetera.

The assumption is that if the subject has complete knowledge of the answer ($x_j=1$), the correct alternative is chosen; that is, $\Phi^{X_jY_j}_{1\,y_j}$ must equal 1 if y_j is the right alternative and 0 if y_j is the wrong alternative. If the subject is in the "Don't know" state ($x_j=0$), $\Phi^{X_jY_j}_{0\,y_j}$ can take on any value from 0 to 1 as long as the sum of the probabilities for all values of y_j (1 through r_j) is 1. The latent responses are assumed to be governed by a one-parameter-logistic model (Rasch, 1960/1980), in which the probability of a latent response x_j, given that the subject has ability θ, is

$$P(x_j|\theta) = \exp(x_j(\theta-\delta_j))/[1 + \exp(\theta-\delta_j)] \tag{2.2}$$

and δ_j is the difficulty of item j.

Assuming that y_j only depends on x_j and that x_j only depends on the latent ability θ, we have

$$P(y_j|\theta) = [\Phi^{X_jY_j}_{0\,y_j} + \Phi^{X_jY_j}_{1\,y_j} \exp(\theta-\delta_j)]/[1+\exp(\theta-\delta_j)]\,. \tag{2.3}$$

In the foregoing, we indicated that an item shows DIF if the probability of a correct response among equally able test takers is different between subgroups. With respect to (2.3), this means that if item j shows DIF, the attraction parameter $\Phi^{X_jY_j}_{x_jy_j}$ and/or the difficulty parameter δ_j did not have the same value for all subgroups. So the two sources of DIF (attractiveness of the alternatives and difficulty of the item) are well-defined by the model.

In order to formulate a complete model, the response pattern of a subject on all the items in a test is denoted by the vector $\mathbf{y}=(y_1,...,y_k)$. The vector of latent responses of a subject is denoted by $\mathbf{x}=(x_1,...,x_k)$. The corresponding random variables are denoted by $\mathbf{Y}$ and $\mathbf{X}$. Furthermore, $F(\theta)$ denotes the continuous distribution function of the latent ability θ, $\delta=(\delta_1,...,\delta_k)$ the vector of item difficulties and $t=x_1+...+x_k$ the number correct score. With the use of (2.1), (2.2), and the assumption of local independence of the y_j and x_j variables, the marginal probability of the observed responses **y** can be written as

$$(2.4) \qquad P(\mathbf{y}) = \sum_{\mathbf{x}} \int_{-\infty}^{\infty} P(\mathbf{y}|\mathbf{x},\theta)\, P(\mathbf{x}|\theta)\, dF(\theta)$$

$$= \sum_{\mathbf{x}} [\prod_{j=1}^{k} \Phi_{x_j y_j}^{X_j Y_j}]\exp(-\sum_{j=1}^{k} x_j\delta_j) \int_{-\infty}^{\infty} \exp(t\theta) C(\theta,\delta)^{-1} dF(\theta) ,$$

in which

$$C(\theta,\delta) = \prod_{j=1}^{k} [1 + \exp(\theta-\delta_j)] ,$$

and $\Sigma_{\mathbf{x}}$ is the summation over all possible latent response patterns $\mathbf{x}=(x_1,...,x_k)$.

In order to detect DIF in multiple-choice items, (2.4) must be extended to include subgroups. In order to keep the main idea of this section in proper perspective, subgroups have been ignored so far but they will be considered in a later section.

In the next section we will formulate the model as an incomplete latent class model. The integral in (2.4) will then be absorbed into a latent class parameter that depends only on the number correct score t, which implies that it is not necessary to specify the distribution function F(θ) any further.

2.4 THE MODEL WRITTEN AS AN INCOMPLETE LCA MODEL

Kelderman (1988) showed that the model in (2.4) is an incomplete latent-class model in the sense of Haberman (1979, chap. 10):

$$(2.5) \qquad P(\mathbf{y}) = \sum_{\mathbf{x}} \Phi_t^T \Phi_{x_1}^{X_1} \ldots \Phi_{x_k}^{X_k} \Phi_{x_1 y_1}^{X_1 Y_1} \ldots \Phi_{x_k y_k}^{X_k Y_k} ,$$

with

$$\Phi_t^T = \int_{-\infty}^{\infty} \exp(t\theta)C(\theta,\delta)^{-1}dF(\theta),$$

and for j=1,...,k,

$$\Phi_{x_j}^{X_j} = \exp(-x_j\delta_j),$$

and in which the attraction parameters are subject to the restrictions

$$(2.6) \qquad \Phi_{x_j 1}^{X_j Y_j} + \dots + \Phi_{x_j r_j}^{X_j Y_j} = 1, \qquad (j=1,\dots,k).$$

In this model, each value of **x** represents a latent class. The model in (2.5) is incomplete, because for certain given values of **X** only a limited number of combinations $(Y_1,\dots,Y_k)$ are possible. Since Φ_t^T depends on an underlying latent trait distribution $F(\theta)$, these parameters are subject to the following complex inequality constraints (Cressie & Holland, 1983; Kelderman, 1984):

$$\det.(\| \Phi_{r+s}^T \|_{r,s=0}^{q_1}) \geq 0,$$

and

$$\det.(\| \Phi_{r+s+1}^T \|_{r,s=0}^{q_2}) \geq 0,$$

in which

$$q_1 = \begin{cases} k/2 & \text{if k is even,} \\ (k-1)/2 & \text{if k is odd,} \end{cases}$$

$$q_2 = \begin{cases} (k-2)/2 & \text{if k is even,} \\ (k-1)/2 & \text{if k is odd,} \end{cases}$$

and $\det.(\| - \|_{r,s=0}^{q})$ defines the determinant of a matrix with row index r and column index s, both running from zero to q.

Since it is not our objective to fit a model for the data, but only to decide if a certain item shows DIF, we will follow Cressie and Holland and ignore these inequality constraints. The resulting model, the so-called generalized Rasch model, provides an easy way to decide whether or not an item shows DIF. The generalized Rasch model is also equivalent to the "conditional" Rasch model; that is, a Rasch model in which there is a conditioning on the number correct score (Kelderman, 1984). The incomplete table methodology can be used to formulate several hypotheses about DIF by specifying alternative models that contain additional subgroup-dependent parameters.

2.5 TESTING FOR DIF BY RELATED LCA MODELS

An item can show DIF in two different ways. First, as indicated before, an item shows DIF if equally able individuals from different subgroups have different probabilities of "Knowing" the answer. This will be referred to as DIF in the latent response. It was assumed earlier that if subjects are in the "Know" state, they will choose the correct alternative. But if subjects are in the "Don't know" state, they may choose any of the alternatives. Therefore, an item also shows DIF if the attractiveness of the alternatives varies from subgroup to subgroup conditioned on ability. This will be referred to as DIF in the attraction parameters or differential alternative functioning (DAF).

In order to detect DIF, the model in (2.5) is reformulated as

$$P(y|i) = \sum_{x} \Phi_{it}^{IT} \Phi_{ix_1}^{IX_1} \ldots \Phi_{ix_k}^{IX_k} \Phi_{ix_1y_1}^{IX_1Y_1} \ldots \Phi_{ix_ky_k}^{IX_kY_k}, \tag{2.7}$$

in which P(y|i) is the conditional distribution of observed response **y** given observed subgroup i (i=1,...,g) and each term on the right side is equal to the corresponding term on the right side of (2.5), extended with the subgroup.

In the model in (2.7), all items are considered to show DIF both in the latent response and the attraction parameters. If some items show DIF neither in the latent response nor in the attraction parameters, the Φ parameters for these items are restricted. If, for example, in a certain model Item 1 shows no DIF in the latent response, the $\Phi_{ix_1}^{IX_1}$ parameters are restricted in the following manner

$$\Phi_{1x_1}^{IX_1} = \ldots = \Phi_{gx_1}^{IX_1} .$$

In the next subsections, models are formulated for the study of the two types of DIF.

2.5.1 DIF in the Latent Response

In order to test whether the interaction between subgroup i and the latent response to Item 1 is zero (i.e. whether Item 1 shows DIF in the latent response), an alternative model is formulated as

$$(2.8) \qquad P(y|i) = \sum_{x} \Phi^{IT}_{it} \Phi^{IX_1}_{ix_1} \Phi^{X_2}_{x_2} \ldots \Phi^{X_k}_{x_k} \Phi^{X_1Y_1}_{x_1y_1} \ldots \Phi^{X_kY_k}_{x_ky_k} .$$

The model in (2.8) can be obtained from (2.7) by setting all Φ-parameters, excluding the difficulty parameter of Item 1, equal for all subgroups:

$$\Phi^{IX_j}_{1x_j} = \ldots = \Phi^{IX_j}_{gx_j} = \Phi^{X_j}_{x_j} , \qquad (j=2,\ldots,k)$$

and

$$\Phi^{IX_jY_j}_{1x_jy_j} = \ldots = \Phi^{IX_jY_j}_{gx_jy_j} = \Phi^{X_jY_j}_{x_jy_j} \qquad (j=1,\ldots,k)$$

This alternative model is compared with

$$(2.9) \qquad P(y|i) = \sum_{x} \Phi^{IT}_{it} \Phi^{X_1}_{x_1} \ldots \Phi^{X_k}_{x_k} \Phi^{X_1Y_1}_{x_1y_1} \ldots \Phi^{X_kY_k}_{x_ky_k} ,$$

in which all Φ-parameters are set equally across subgroups. If a statistical test of the difference between the models is significant, we may conclude that the difficulty of Item 1 varies from subgroup to subgroup. In this case, subjects in one subgroup may find it more difficult to solve the problem than subjects in another subgroup.

2.5.2 DIF in the Attraction Parameters

In order to test the null hypothesis that the interaction between the subgroup and the observed response to Item 1 is zero (i.e. whether Item 1 shows DIF in the attraction parameters), (2.9) is compared with the alternative model

$$(2.10) \qquad P(y|i) = \sum_{x} \Phi^{IT}_{it} \Phi^{X_1}_{x_1} \ldots \Phi^{X_k}_{x_k} \Phi^{IX_1Y_1}_{ix_1y_1} \Phi^{X_2Y_2}_{x_2y_2} \ldots \Phi^{X_kY_k}_{x_ky_k} ,$$

in which, similar to (2.8), all Φ-parameters, except for the attraction parameters for Item 1, are set equal across subgroups. If the statistical test is significant, we may conclude that the attractiveness of the Item 1 alternatives varies from subgroup to subgroup. In (2.8) and (2.10), the Φ-parameters are specified to test for DIF for only one item. Obviously, similar model terms can be specified for two or more items if necessary. It is also possible to define models in which one item shows DIF in the latent response and another (or the same) item shows DIF in the attraction parameters.

2.6 PARAMETER ESTIMATION AND MODEL TESTING

Let n_{ixy} be the number of individuals in subgroup i with X=x and Y=y under a certain model and let m_{ixy} be the expected value of n_{ixy}. Although n_{ixy} is not observed, it is possible to estimate the expected value m_{ixy} of n_{ixy}, and the Φ-parameters from the observed n_{iy} (or n_y if the subgroup is unobserved) by the method of maximum likelihood. To illustrate this, consider the model in (2.7). The likelihood equations for (2.7) would be (Haberman, 1979):

$$m^{IT}_{it} = \hat{n}^{IT}_{it}, \quad m^{IX_jY_j}_{ix_jy_j} = \hat{n}^{IX_jY_j}_{ix_jy_j}, \qquad (j=1,\ldots,k),$$

in which

$$\hat{n}_{ixy} = (m_{ixy}/m_{iy})\, n_{iy}, \tag{2.11}$$

and n^{IT}_{it} and $n^{IX_jY_j}_{ix_jy_j}$ are the numbers of individuals in subgroup i with T=t, $X_j=x_j$, and $Y_j=y_j$, respectively. Furthermore, m^{IT}_{it} and $m^{IX_jY_j}_{ix_jy_j}$ are the expected values of n^{IT}_{it} and $n^{IX_jY_j}_{ix_jy_j}$, respectively. If the subgroup i is not observed, n_{iy} and m_{iy} in (2.11) have to be replaced by n_y and m_y, respectively. The likelihood equations can be solved by the iterative proportional fitting algorithm or the scoring algorithm (Goodman, 1978; Haberman, 1979). The iterative proportional fitting algorithm is preferred, because it is less sensitive to the choice of starting values. Similar likelihood equations can be formulated for the restricted models.

The overall goodness-of-fit of an incomplete latent-class model can be tested by the Pearson statistic (Q) or the likelihood-ratio statistic (LR) (see Haberman, 1979). Both statistics are asymptotically distributed as chi-square with degrees of freedom equal to the difference between the number of possible response patterns **y** (or this number multiplied by g if the subgroup is observed) minus one and the number of estimable parameters. The number of estimable parameters of a model should be equal to the rank of the information matrix (Goodman, 1978; McHugh, 1956).

By the difference in the likelihood ratio test statistics for two models (LR(a;b)), it can be tested whether the alternative model *b* yields a significant improvement in fit over the restricted model *a*, which is a special case of model *b*. Under the assumption of model *a*, LR(a;b) is asymptotically chi-square distributed with degrees of freedom equal to the difference in the numbers of estimable parameters in both models (Bishop, Fienberg, & Holland, 1975).

2.7 AN EMPIRICAL EXAMPLE

As an example, four items from the Second International Mathematics Study in the Netherlands were analyzed (Eggen, Pelgrum, & Plomp, 1987). Each item was a five-choice item with only one correct alternative. A sample of 3002 students was drawn from two types of schools for lower secondary education in the Netherlands. To illustrate the use of quasi-loglinear models for the detection of DIF, the students' level of education was chosen as the grouping variable: subgroup MAVO (intermediate general education) and subgroup HAVO/VWO (higher general education and pre-university education).

The models in (2.8) and (2.10) were fitted to the data with the computer-program LCAG (Hagenaars & Luijkx, 1990). LCAG is a program for the estimation of the parameters of loglinear models with latent variables, and yields, beside the estimated latent conditional probabilities (i.e. the attraction parameters), the estimated expected frequency distribution of the latent variables within the model. From this frequency distribution the difficulty parameters were estimated through LOGIMO (Kelderman & Steen, 1988). LOGIMO is a general computer program especially written for the analyzation of loglinear IRT models. In both programs the efficient IPF algorithm is used for the estimation of the parameters.

In the first series of analyses, each item was separately tested for DIF in the latent response or in the attraction parameters. For example, to test if Item 1 showed DIF in the latent response, we postulated that the difficulty parameter of Item 1 was the only parameter that varied between the two groups. The models in (2.8) and (2.10) were compared to (2.9) to test for DIF in the latent response and for DIF in the attraction parameters, respectively. Table 2.1 shows the values of the likelihood ratio test and the degrees of freedom for the models in (2.8) and (2.10), for each item separately. In both tests, group membership (i.e. the level of education) was assumed to be known.

Table 2.1

Likelihood-ratio tests statistics for the detection of DIF in the data from the Second International Mathematics Study

	DIF in the latent response		DIF in the attraction parameters	
Item	Likelihood-ratio	df	Likelihood-ratio	df
1	1.701	1	26.519*	4
2	4.720*	1	21.340*	4
3	1.747	1	6.033	4
4	.018	1	52.595*	4

Note: Tests marked with an asterisk are significant ($\alpha = .05$).

From Table 2.1 we may conclude that, except for Item 2, the item difficulty parameters do not vary significantly between the two subgroups (MAVO and HAVO/VWO). Only Item 2 shows DIF in the latent response. When we take a closer look at the difficulty of Item 2, we can see that it was substantially smaller for MAVO-students ($\delta_{22} = 1.90$) than for HAVO/VWO-students ($\delta_{12} = 0.82$). The difficulty parameters of the other three Items 1, 3, and 4 were -1.52, -3.54, and 1.32, respectively. Please note that these items showed no DIF in the latent response; therefore, the difficulty parameters were estimated by setting the item parameters equal in both subgroups.

The test LR(2.9;2.10) reported in Table 2.1 also indicates that the attractiveness of the alternatives to Items 1, 2, and 4 were significantly different for both subgroups. Estimates of the attraction parameters for the alternatives of each item are presented in Table 2.2. These results indicate that a HAVO/VWO-student is more likely to choose the correct alternative to Item 1 than a MAVO-student. On the other hand, a MAVO-student is more likely to choose the correct alternative for Item 2, because the associated attraction parameter of the correct alternative for Item 2 in this group is twice as large as the associated attraction parameter for a HAVO/VWO-student. For both subgroups, however, the correct alternative is not the most attractive one.

The attraction parameters for the correct alternative of Item 4 are approximately the same for both subgroups, but for the alternatives B and C, a curious difference exists between the two subgroups. A HAVO/VWO-student would choose alternative B with almost the same probability as a MAVO-student would guess alternative C, and (s)he would choose alternative C with almost the same probability as a MAVO-student would choose alternative B. Item 3 shows no DIF in the attraction parameters. However, the attraction parameter for the right alternative in the subgroup HAVO/VWO is more than three times as large as the associated attraction parameter

Table 2.2
Attraction parameters for the alternatives of the four items

	Alternatives Subgroup HAVO/VWO					Alternatives Subgroup MAVO				
Item	A	B	C	D	E	A	B	C	D	E
1	.073	.033	.685	.174	.035	.211	.024	.563	.193	.009
2	.743	.123	.061	.045	.028	.662	.240	.068	.015	.015
3	.106	.296	.146	.367	.085	.140	.468	.122	.111	.159
4	.110	.355	.235	.092	.208	.068	.241	.341	.084	.266

Note: The correct alternatives are underlined.

in the subgroup MAVO. Nevertheless, this difference had no significant effect on the test for DIF in the attraction parameters, because the item was very easy for both subgroups.

A major problem in DIF studies is the explanation of DIF when it is observed. Although it is beyond the scope of this chapter, a tentative explanation for the observed DIF in the attraction parameters of Item 4 is the subjects familiarity with the mathematical terms. In Item 4 (see Appendix A.1) the subject is asked to give the definition of a parallelogram. Since the attraction parameters for the alternatives A, D, and E are approximately the same for the two subgroups (see Table 2.2), the observed DIF in the attraction parameters is probably caused by the alternatives B and C. Knowing the formulation of Item 4 we can conclude that a MAVO student is probably more familiar with the mathematical terms axis of symmetry and diagonal than a HAVO/VWO student.

In the foregoing analyses the two types of DIF were studied separately. Moreover, only one item was evaluated at a time. As indicated earlier, it is possible to analyze models in which more than one item shows DIF. In order to illustrate this possibility, a model was considered in which Items 1, 2, and 4 show DIF in the attraction parameters and Item 2 shows DIF in the latent response. This model shows considerable improvement in fit, compared to the model in (2.9) (likelihood-ratio is 100.5 with 13 degrees of freedom). From Table 2.1 it also follows that this model fits the data better than the models previously discussed. The estimated parameters, however, do not differ much from the estimated parameters of the previous models; therefore, they are not given.

In summary, the difficulty of the four items can be ordered in the following way: $\delta_3 < \delta_1 < \delta_4 < \delta_2$. That is, Item 3 is the easiest and Item 2 is the most difficult. The attractiveness of alternatives of Items 1, 2, and 4 as well as the difficulty of Item 2 are not the same for the two subgroups. Item 3 shows no DIF in the latent response or in the attraction parameters.

2.8 DISCUSSION

In this chapter we proposed an incomplete latent class model for the examination of DIF in multiple-choice items through a combination of the usual notion of DIF for correct/incorrect responses and information about DIF incorporated by each of the alternatives. In the method proposed a distinction is made between a "Know" state in which the subject has a complete knowledge of the answer and a "Don't know" state. It is assumed that if the subject is in the "Know" state, (s)he will give the correct answer. The probability that the subject is in the "Know" state is assumed to be governed by the Rasch model. And, if the subject is in the "Don't Know" state, the subject will choose the most attractive alternative, in which the attractiveness of the alternatives may be different for different alternatives, including the correct one. In order to study DIF the model is extended with variables (observed or latent) which determine subgroup membership. One of the main advantages of the proposed method is that it is not only possible to test if a certain item shows DIF, but it is also possible to test whether this DIF is caused by the difficulty of the item, the attractiveness of the alternatives, or both.

In most applications, the subgroup membership is determined by an observed variable (e.g. sex). In some situations, however, subgrouping is suspected but the variable determining subgroup membership cannot be observed (Kelderman & Macready, 1990; Mislevy & Verhelst, 1990). When no subgroup membership can be established, the subgroup variable in the proposed method is also treated as a latent variable.

In this chapter all tests of DIF are two-sided. This means that it is not possible to test directional hypotheses about DIF. The estimated difficulty parameters and the estimated attraction parameters only give an indication of the direction of DIF. However, together with the knowledge of the item, these estimated parameters may provide the test-constructor a better feel for the reason why an item does or does not show DIF. Furthermore, if many items in a test show DIF, it might be that DIF in one of the items in favor of a subgroup is compensated by DIF in another item in favor of another subgroup. And, although DIF in the attraction parameters may have no effect on test scores, it could indicate that the item was functioning differently for the different subgroups.

At the present time the method is not very practical for a large number of polytomous items. This problem is due to the computer program LCAG, in which in our case such a large amount of memory space is required that it is impossible to consider more than four five-choice items at a time. A line of future research should be the development of an estimation method which can handle many items.

Chapter 3

THE ESTIMATION OF THE PARAMETERS OF THE SOLUTION-ERROR RESPONSE-ERROR MODEL WITH THE USE OF SUBSETS OF ITEMS

3.1 ABSTRACT

In Westers and Kelderman (1992) the solution-error response-error model is formulated as a latent class model for the incomplete subgroup x item response 1 x...x item response k contingency table. Parameter estimates can be computed with the programs LCAG and LOGIMO, but this becomes unpractical if the number of items is large. In that case the tables of observed and expected counts become too large for computer storage in LCAG.

In this chapter a method of parameter estimation is described that is based on the division of the entire item set into several subsets. The computational problem boils down to the estimation of the parameters of a set of smaller solution-error response-error models. It is shown that, depending on the number of items in each subset, the total number of cells in the tables of observed and expected counts can be considerably reduced by this method. In this way, models with a large item set may be computed much more efficiently, in terms of both computer storage and processing time.

3.2 INTRODUCTION

In the solution-error response-error (SERE) model (Kelderman, 1988), a distinction is made between a "Know" state, in which the subject has complete knowledge of the answer, and a "Don't know" state. The probability that the subject is in the "Know" state is assumed to be governed by the Rasch (1960/1980) model. If the subject is in the "Don't know" state, the subject will guess the most attractive alternative, where the attractiveness of an alternative may be dissimilar for different alternatives, including the correct one. The SERE model can easily be formulated as a latent class analysis model (Kelderman, 1988) in which the structure of the latent-class probabilities is explained by a loglinear Rasch model. In the SERE model, each latent class corresponds with an idealized response pattern. The relations between these idealized responses are explained by the loglinear version of the Rasch model (Kelderman, 1984). Parameter estimates can be computed with the programs LCAG (Hagenaars & Luijkx, 1990) and LOGIMO (Kelderman & Steen, 1988), but the software becomes unpractical if the number of items is large (Westers & Kelderman, 1992). In the first place, in LCAG all cell frequencies, including empty cells, have to be listed. Secondly, for each latent class both the probability for the latent class and the conditional probabilities of the observed variables given the latent class, have to be given in LCAG. By fulfilling these two requirements, LCAG requires such a large amount of memory space that it is impossible to consider a large item set. Therefore, in this chapter an alternative estimation method will be proposed for the SERE model that can handle a larger set of items. In the remainder of this chapter, the proposed estimation method will be described, but the maximum likelihood estimation methods that are currently in use will be discussed first.

Latent class analysis (LCA) models have been used extensively for measurements in sociology, psychology, and education (Clogg, 1981). There is a well-developed theory of maximum likelihood estimation and likelihood-ratio testing of LCA models. McHugh (1956) derived the maximum likelihood estimators, but his solution applies only to the unconstrained model. Great progress was made when Goodman (1974b) described a particularly simple iterative procedure which also has the virtue of automatically producing estimates of probabilities that always fall in the unit interval. Furthermore, it is very easy to modify the procedure to satisfy a reasonable variety of other constraints on the parameters. This simple iterative procedure is used in the program MLLSA (Clogg, 1977; Eliason, 1988), LCAG (Hagenaars & Luijkx, 1990), PANMARK (van de Pol, Langeheine, & de Jong, 1989) and MIRA (Rost & von Davier, 1992). The latter two programs are developed for some extensions of latent class analysis that violate the basic assumption of local independence: the mixed latent Markov model of Langeheine and van de Pol (1990; van de Pol & Langeheine, 1990) and the mixed (polychotomous) Rasch model of Rost (1990, 1991).

Another estimation procedure based on the maximum likelihood principle is the method of marginal maximum likelihood (MML). In this method the assumption is that the subjects are drawn at random from a population of abilities. For the IRT model, the method of MML was first applied by Bock and Aitkin (1981) and Thissen (1982). They used two methods of solving the marginal likelihood equations: the so-called EM method and Newton-Raphson iterations. In the paper by Paulson (1986), a review is given of the application of the EM algorithm of

Dempster, Laird and Rubin (1977) to MML estimation of parameters in the LCA model. Paulson also shows how the EM-algorithm can be used to obtain marginal maximum likelihood estimates of the item response functions under the minimal monotone homogeneity assumption. To avoid the unmanageability of the contingency table as the number of items increases, Paulson's algorithm deals with each individual response vector, rather than cell counts in a item 1 by item 2 by ... item k contingency table. Therefore, the effect of increasing the number of items has no effect on the algorithm beyond the increase of running time, which is directly proportional to the number of items (Paulson, 1986).

Under certain item response models (Rasch, 1960/1980) it is possible, by conditioning on the number correct-scores (i.e. the sufficient statistics), to get a conditional likelihood that only depends on the item parameters. Conditional maximum likelihood (CML) estimation proceeds under such models by maximizing this conditional likelihood. The advantages of CML estimation are that the estimators of the item parameters are consistent and that the well-known theorem on the asymptotic normality of ML estimators holds (Andersen, 1973). The disadvantages of CML estimation are that this estimation method is only possible for the Rasch model (Thissen, 1982) and that some information about the item parameters in the data is disregarded. For a long time, CML estimation has only been possible for a small number of items (Hambleton & Swaminathan, 1985), but Verhelst, Glas and van der Sluis (1984) and Verhelst and Veldhuijzen (1991) have shown that as many as a thousand items can now be dealt with.

Although ML estimation has many appealing statistical properties, other good estimation procedures are available. The methods based on minimum chi-square (MCS) is one of these competing estimation methods. In the MCS procedures the data are grouped into mutually exclusive and exhaustive classes, and distance functions of the observed and expected frequencies in these classes are defined. Minimalization of these distance functions provide the parameter estimators of the model. Some well-known examples of MCS procedures are the Pearson chi-square, the likelihood chi-square and Neyman's reduced chi-square method. Most minimum chi-square (MCS) estimates are much easier to evaluate than ML estimates (Cramér, 1946; Engelen, 1989), although the simplifications are only slight. For LCA models it can be shown that MCS procedures fall into the multinomial case (Engelen, 1989); the only difference with the ordinary Rasch model is that there are now latent classes of examinees instead of individual examinees.

Furthermore, McHugh (1956, Mooijaart, 1978) showed that the latent class model met the requirements of the general theorem of Neyman (1949, p. 250), which implies that the estimators obtained as the solution of the likelihood equations are, in fact, asymptotically normally distributed, and are the "best" estimators in the sense that they have the smallest covariance matrix. Such estimators are called the best asymptotically normal (BAN) estimators. Both the maximum likelihood estimator and the minimum chi-square estimator are examples of BAN estimators.

The well-known algorithms of Goodman (1974a, 1974b) and Haberman (1979) for the estimation of the parameters of loglinear models with latent classes (e.g. the SERE model) can only be used for a small number of variables. For models with a large number of variables these algorithms become very complex. For the general loglinear model an algorithm has been developed which is based on a faster way of calculating the sufficient statistics of the parameters in the model; the so-called marginalization-by-variable principle (Kelderman, 1992). In this chapter another principle will be introduced, which has the advantage that for a large number of items the parameters of the model can still be estimated by the well-known EM algorithm of Dempster, Laird and Rubin (1977). The main idea of this new principle is to divide the entire item set into several subsets. By doing this, the SERE model can be rewritten into a set of smaller SERE models. By estimating the parameters of these smaller SERE models simultaneously, it is possible to estimate the item parameters of the entire SERE model. A similar approach has been developed by Mellenbergh and Vijn (1981) for the estimation of the parameters in the Rasch model. Instead of the full item 1 x...x item k x sum-score table, they studied the item response x sum-score tables for each item.

One of the main advantages of our approach is the decreased total number of cells in the marginal contingency tables, especially when there are many items. A second advantage is the decreased memory space needed to store information about the latent classes. In the third place, the proposed procedure permits a practical use of (incomplete) response data. A disadvantage is that some of the statistical efficiency of the estimators may be lost when the SERE model is collapsed.

Below the solution-error response-error (SERE) model for polytomous items will be developed and formulated as an LCA model. A new computationally efficient estimation method for the SERE model for large sets of polytomous items is described and its use is illustrated by means of simulated data.

3.3 THE SOLUTION-ERROR RESPONSE-ERROR MODEL

Suppose that each subject, randomly drawn from a population of subjects, responds to k test items, where the answer to item j may be any of the r_j responses, denoted by y_j ($y_j=1,\dots,r_j$). Let x_j indicate the latent response of the subject. The assumption is that the latent responses are governed by a one-parameter-logistic model (Rasch, 1960/1980), in which the probability of the latent response x_j (x_j=0,1), given that the subject has ability θ, is

(3.1) $$P(x_j|\theta) = \exp(x_j(\theta-\delta_j))/[1+\exp(\theta-\delta_j)]$$

and δ_j is the difficulty of item j.

The relationship between the latent response x_j and the observed response y_j is described by the conditional probability

(3.2) $$\Phi^{X_jY_j}_{x_jy_j} \equiv P(y_j|x_j) \, ,$$

in which the superscripts are symbolic notations that indicate that the random variables X_j and Y_j are involved in the definition of the conditional probability. For the sake of simplicity, the notation y_j, x_j, etc. in the probabilities is used for Y_j=y_j, X_j=x_j, et cetera.

To formulate a complete model, the response pattern of a subject on all k items in a test is denoted by the vector $\mathbf{y}=(y_1,...,y_k)$. The vector of latent responses of a subject is denoted by $\mathbf{x}=(x_1,...,x_k)$. The corresponding random variables are denoted by **Y** and **X**. Let $F(\theta)$ be the continuous distribution function of the latent ability θ, $\delta=(\delta_1,...,\delta_k)$ and $t=x_1+...+x_k$ the number correct score. With the use of (3.1), (3.2) and the assumptions that y_j depends only on x_j, and x_j depends only on the latent ability level θ, Kelderman (1988) has shown that the marginal probability of the observed responses **y** can be written as a latent-class model in the sense of Haberman (1979, chap. 10). If $\Sigma_{\mathbf{x}}$ is the summation over all possible latent response patterns $\mathbf{x}=(x_1,...,x_k)$, then

(3.3) $$P(\mathbf{y}) = \sum_{\mathbf{x}} \Phi^{T}_{t} \, \Phi^{X_1}_{x_1} \cdots \Phi^{X_k}_{x_k} \, \Phi^{X_1Y_1}_{x_1y_1} \cdots \Phi^{X_kY_k}_{x_ky_k} \, ,$$

with

$$\Phi^{T}_{t} = \int_{-\infty}^{\infty} \exp(t\theta)C(\theta,\delta)^{-1} dF(\theta) \, ,$$

$$C(\theta,\delta) \equiv \prod_{j=1}^{k} [1+\exp(\theta-\delta_j)] \, ,$$

and for j=1,...,k,

$$\Phi^{X_j}_{x_j} = \exp(-x_j\delta_j)$$

and in which the attraction parameters (Φ^{XY}) are subject to the restrictions

$$\Phi^{X_jY_j}_{x_j 1} + \dots + \Phi^{X_jY_j}_{x_j r_j} = 1\,, \qquad (j=1,\dots,k). \tag{3.4}$$

In this model, each value of the latent response vector **x** represents a latent class. If certain conditional probabilities $P(y_j|x_j)$ are specified to be zero, the model in (3.3) is incomplete, because in that case, for certain given values of **X**, not all combinations of **Y** are possible.

Since the sum-score parameters (Φ^T) depend on the underlying distribution of the latent scores, they are subject to complex inequality restrictions (Cressie & Holland, 1983). There are, however, no restrictions on the sum-score parameters for the conditional Rasch model; that is, a Rasch model conditioned on the number correct score (Kelderman, 1984). Throughout this chapter we either assume that these restrictions hold or we work with the conditional model.

Furthermore, the multiplication of each difficulty parameter (Φ^X) by a constant c and the division of each sum-score parameter (Φ^T) by c^t, does not change the model in (3.3). This indeterminacy can be removed by setting one of the item difficulties (δ_j) equal to zero.

3.3.1 Restrictions on the model parameters

In the same way as Goodman (1974b) formulated restrictions on the parameters of the LCA models, restrictions may occur with respect to the parameters of the SERE model. In the first place, the attraction parameters may be equated with each other or with a prespecified value. Relevant constraints are, especially, those that set the attraction parameters to the values 0 or 1 to indicate, for example in the case of multiple choice items, that a subject will choose the right alternative if the subject is in the "Know" state (Westers & Kelderman, 1992). Secondly, equality restrictions may be used to make the alternatives equally attractive. Like the attraction parameters, the difficulty parameters may also be equated with each other or with a constant (including 0). Finally, equality restrictions on the attraction parameters or difficulty parameters may be used to examine differential item functioning (DIF) in polytomous or dichotomous items, as discussed by Westers and Kelderman (1992).

For the general case of equality constraints, Mooijaart and van der Heijden (1992) have shown that "... the EM-algorithm is not simple to apply because a nonlinear equation has to be solved. This problem arrives, mainly, when equality constraints are defined over probabilities in different combinations of variables and latent classes" (p. 261). Mooijaart and van der Heijden have given a simple condition in which, although the restriction remains that the probabilities in different variable-latent class combinations are equal, the EM-algorithm is still simple to apply. In words their condition reads: "(1) In cases where each of the equality constraints holds only for the parameters in one variable-latent class combination, the standard EM algorithm estimation procedure gives correct results; (2) In cases where the number of elements of an equality set is equal for different variable-latent class combinations, the standard EM algorithm estimation procedure is correct, assuming that the fixed elements are zero. When the fixed elements are non-zero, the condition is more complicated; (3) In all other cases, for each EM step, estimation of the parameters has to be done by an iterative procedure" (p. 268).

The maximum likelihood estimates of the parameters of the model in (3.3) can then be obtained by solving the likelihood equations by the iterative proportional fitting (IPF) algorithm. Computer programs by Hagenaars and Luijkx (LCAG, 1990) and Kelderman and Steen (LOGIMO, 1988) can be used to fit the model. The overall goodness-of-fit of a model can be tested by the Pearson statistic or the likelihood-ratio test statistic (see Haberman, 1979). In the next sections, we shall introduce a new method for obtaining the maximum likelihood estimates of the parameters in the SERE model.

3.4 THE DIVISION-BY-ITEMS PRINCIPLE IN THE SOLUTION-ERROR RESPONSE-ERROR MODEL

As already noted by Westers and Kelderman (1992), the model in (3.3) is only usable in practice when the responses to few items are studied. One of the solutions to this problem could be not to consider all items simultaneously. In this section a new estimation method is proposed which is based on the division of the entire item set into several subsets. We will refer to this operation as an application of the division-by-items (DBI) principle. In this section the new estimation method is explained for pairs of items. It may be clear that the results in this section would not change if we consider subsets of three or more items or subsets with unequal numbers of items.

Let $P(y_1,y_2)$ be the probability that the observed response on item 1 is y_1 and the observed response on item 2 is y_2. If we let $z=x_1+x_2$ and use conditional probability calculus and elementary calculus, it can be shown that the model in (3.5) is also a latent class model (see Appendix A.2)

$$P(y_1,y_2) = \sum_{y_3} \ldots \sum_{y_k} P(y_1,\ldots,y_k) = \sum_{y_3} \ldots \sum_{y_k} P(y) \tag{3.5}$$

$$= \sum_{x_1} \sum_{x_2} \Phi_z^{T12} \; \Phi_{x_1}^{X_1} \; \Phi_{x_2}^{X_2} \; \Phi_{x_1 y_1}^{X_1 Y_1} \; \Phi_{x_2 y_2}^{X_2 Y_2},$$

in which

$$\Phi_z^{T12} = \int \exp(z\theta) \, \{[1+\exp(\theta-\delta_1)][1+\exp(\theta-\delta_2)]\}^{-1} \, dF(\theta),$$

which is similar to (3.3), except that here we consider two items and in (3.3) k items. This means that, given the assumption of local independence, the SERE model is collapsible in the sense that taking the marginal probability for two items from the entire SERE model yields the SERE model for two items. We will refer to these smaller SERE models as the collapsed SERE models. The way in which consistent and asymptotic normal estimators for the parameters of the SERE model can be obtained from the maximum likelihood estimators of the collapsed SERE models will be discussed in Section 3.5. However, since information about the joint relationships among items may be lost when the SERE model is collapsed, these estimators will not be efficient. Maximum likelihood estimates of the parameters of each collapsed SERE model can be obtained by solving the likelihood equations by the iterative proportional fitting (IPF) algorithm.

In order to obtain the same measurements from different subsets of items, the subsets must measure the same ability and the scores must be measured on the same scale. In that case the subsets are said to be equated. Generally, subsets of items can be equated on the same scale if each subset is directly or indirectly connected to all other subsets by common items (Wright, 1977). Fischer (1974, 1981) has shown that unique (conditional) maximum likelihood estimates exist in the Rasch model, if and only if in every possible partition of the items into two (nonempty) subsets, some subject has responded correctly to some item from the first subset and responsed incorrectly to some item from the second subset. He has even generalized this conditional for the case of the polytomous multidimensional Rasch model.

With respect to the condition of Wright, the division of the set of k items into a set of non-overlapping subsets of items would not give the same (conditional) maximum likelihood estimator in each subset. However, dividing the set of k items into the subsets (1,2), (2,3),...,(k-1,k) would meet this condition. On the other hand, the division of the set of items in all possible pairs of items would also meet this condition. There are many more ways of dividing the set of k items into appropriate subsets. An important question is: what is the best selection of appropriate subsets? To answer this question, criteria for an optimum division have yet to be constructed.

Therefore, selecting the optimum division of the set of items into subsets should be a line of future research. In this chapter we will use all possible pairs of items as subsets of items.

Since the subsets of items must not be distinct, the parameters of the collapsed SERE models have to be estimated simultaneously. For example, the parameters (i.e. the attractiveness of the alternatives and the difficulty) of an item in one subset have to be equal to the parameters of the same item in other subsets. With pseudo-likelihood estimates this requirement can be met.

For the present model, a pseudo-loglikelihood can be expressed as the sum of the loglikelihoods for subsets of items. A statistic maximizing a pseudo-likelihood will be termed a pseudo-likelihood estimator. To prove the consistency and asymptotic normality of the maximum pseudo-likelihood estimator, modified classical methods can be used (Arnold & Strauss, 1988). In the next section the pseudo-likelihood theory will be discussed in more detail.

One of the main advantages of the proposed estimation method is the decrease of the total number of cells in the marginal contingency tables, especially when there are many items. Most of the currently available algorithms require the storage of the full observed and expected frequency tables. For example, if there are ten five-response items, each table will consist of about 10 million cells, whereas, if the DBI principle is used, the number of cells equals to the sum of the number of cells in the marginal frequency tables over all subsets. In Table 3.1 the

Table 3.1
Total number of cells in the observed contingency tables

Number of items	Number of alternatives	Number of items in each subset of items				
		2	3	4	5	all
6	2	60	160	240	192	64
	3	135	540	1215	1458	729
	4	240	1280	3840	6144	4096
7	2	84	280	560	672	128
	3	189	945	2835	5103	2187
	4	336	2240	8960	21504	16384
8	2	112	448	1120	1792	256
	3	252	1512	5670	13608	6561
	4	448	3584	17920	57344	65536
9	2	144	672	2016	4032	512
	3	324	2268	10206	30618	19683
	4	576	5376	32256	129024	262144
10	2	180	960	3360	8064	1024
	3	405	3240	17010	61236	59049
	4	720	7680	53760	258048	1048576

total numbers of cells of the observed contingency tables are given for six to ten polytomous items and for different numbers of items in each subset. It can be seen that for the proposed estimation method these numbers remain within reasonable limits, especially for small subsets of items, whereas for the currently available algorithms these numbers increase very rapidly.

Another advantage is associated with the storage of the latent class probabilities P(**x**) and the conditional probabilities $P(y_j|\mathbf{x})$. Most of the currently available algorithms require the storage of all these probabilities. This means that for the SERE model we have to store the probabilities P(**x**) and $P(y_j|\mathbf{x})$ for all 2^k latent classes. Calculations dealing with all these latent classes become impractical very quickly as the number of items increases.

Finally, the proposed procedure permits a practical use of response data. Apart from designs with complete data, designs with incomplete data can also be used. Data from any subject, even when they respond to only two items of the test, can be used in the estimation of the attraction parameters and the difficulty parameters. Of course, the data for these subjects can only be used if the two items form one of the subsets in the proposed estimation method.

A disadvantage of the DBI principle is that information about the joint relationships among the items may be lost when the SERE model is collapsed.

3.5 THE ESTIMATION OF THE PARAMETERS OF THE SOLUTION-ERROR RESPONSE-ERROR MODEL

The SERE model can be seen as an LCA model with a non-saturated loglinear model (i.e. the Rasch model) imposed upon the distribution of the latent classes. Therefore, the maximum likelihood estimates of the parameters of the (collapsed) SERE model can be obtained by solving the likelihood equations by a two-step algorithm. Let us first assume that the latent response vector **x** was observed in addition to the observed response **y**. Then the solution would be at hand; if the responses **x** and **y** are known, the maximum likelihood estimates of the parameters of the SERE model can be found by the usual maximum likelihood methods for the estimation of the parameters of a model (Haberman, 1979; Hagenaars, 1988; Kelderman, 1988). However, the latent responses **x** are not observed, but they can be estimated from the estimated values of the parameters of the SERE model. In the literature this method is known as the EM-algorithm (Dempster, Laird & Rubin, 1977). By the E-step the expectation of the estimated observed number of subjects with latent response **x** and observed response **y** given the number of subjects with observed response **y** can be found, and by the M-step the Φ-parameters can be estimated with complete data maximum likelihood techniques (Rubin, 1991).

In the following sections we will describe the traditional estimation method for the SERE model as discussed in Kelderman (1988) and Westers and Kelderman (1992), discuss the use of the pseudo-likelihood theory in the case of the SERE model, and describe an estimation method based on the pseudo-likelihood theory. Finally, the issue of the initial values and some indices for testing whether an item shows DIF will be discussed.

3.5.1 Traditional estimation method for the SERE model

In the following treatment of the traditional estimation method, we will omit the superscripts in the notation of the parameters.

Let $m_{i\mathbf{xy}} = n_i \, P(\mathbf{x},\mathbf{y}|i)$ be the expected number of subjects in subgroup i (i=1,....,g) with latent response vector **x** and observed response vector **y** under the SERE model (3.3)

$$P(\mathbf{x},\mathbf{y}|i) = \Phi_{it} \, \Phi_{ix_1} \cdots \Phi_{ix_k} \, \Phi_{ix_1y_1} \cdots \Phi_{ix_ky_k} \, , \tag{3.6}$$

in which $P(\mathbf{x},\mathbf{y}|i)$ is the conditional distribution of latent response **x** and observed response **y** given observed subgroup i (i=1,....,g), n_i is the number of subjects in group i and each factor on the right-hand side corresponds with a factor on the right-hand side of (3.3) extended with a variable i for group membership. If we assume that for a particular item j the conditional probability (3.2) is zero for a certain combination of the latent response X_j and observed response Y_j, then the probability (3.6) equals zero.

Let $n_{i\mathbf{xy}}$ denote the unobserved number of subjects in subgroup i (i=1,....,g) with latent response vector **x** and observed response vector **y**. If the latent response vector **x** were observed, the likelihood equations for the model in (3.6) would then be (Haberman, 1979; Hagenaars, 1990)

$$m_{ix_jy_j} = n_{ix_jy_j} \, , \qquad (j=1,....,k), \tag{3.7}$$

$$m_{it} = n_{it} \, , \tag{3.8}$$

in which m_{it} and n_{it} are the expected and observed number of subjects in subgroup i with sum-score t.

However, the response vector **x** is latent and its scores cannot be directly observed. Consequently, n_{it} and $n_{i\mathbf{xy}}$ are not observed. Haberman (1979) has indicated that for LCA models "the same likelihood equations apply as in the ordinary case in which all frequency counts are directly observed, except that the unobserved counts are replaced by their estimated conditional expected values given the observed marginal tables" (p. 543). Let f_{ixy} be the estimated observed number of subjects in subgroup i with latent response $\mathbf{x}$ and observed response y given the observed marginal counts $n_{iy}=\Sigma_{\mathbf{x}}\, n_{ixy}$ given by

$$f_{ixy} = m_{ixy}\, n_{iy} / m_{iy} \ , \tag{3.9}$$

in which $m_{iy}=\Sigma_{\mathbf{x}}\, m_{ixy}$, and let f_{it} be the estimated observed number of subjects in subgroup i with sum-score t defined as

$$\begin{aligned} f_{it} &= \sum_{\mathbf{x}|t} \sum_{y} f_{ixy} \\ &= \sum_{\mathbf{x}|t} \sum_{y} m_{ixy}\, n_{iy} / m_{iy} \ , \end{aligned} \tag{3.10}$$

in which $\Sigma_{\mathbf{x}|t}$ is the summation over all possible latent response patterns $\mathbf{x}$ with sum-score t. The likelihood equations 3.7 and 3.8 can then be replaced by (Haberman, 1979)

$$m_{ix_jy_j} = f_{ix_jy_j} \ , \qquad (j=1,\dots,k), \tag{3.11}$$

$$m_{it} = f_{it} \ . \tag{3.12}$$

Since the SERE model can be transformed into a latent class model, we can use Clogg and Goodman's (1985) extension of Goodman's (1974a, 1974b) variant of the EM-algorithm for solving the likelihood equations. The algorithm works as follows.

First initial values for the parameters on the right-hand side of equation 3.6 are determined. How the initial values can be determined will be discussed in Section 3.5.4. In view of the chosen initial values for the parameter, $m_{i\mathbf{x}y}$ is computed from (3.6) by

$$m_{i\mathbf{x}y} = n_i\, P(\mathbf{x},y|i) \ , \tag{3.13}$$

in which n_i is the number of subjects in group i. The estimated observed frequencies $f_{i\mathbf{x}y}$ are then computed by means of Equation 3.9. This is the Expectation step (E-step) of the EM-algorithm. Next, in the Maximization step (M-step), the model parameters are obtained by solving the Equations 3.11 and 3.12 for the model parameters.

Since we assumed that the observed responses **y** only depend on the latent responses **x** and the latent responses **x** only depend on the ability level θ, the attraction parameters and the difficulty parameters can be computed separately. This can be seen when the joint distribution of the latent responses given subgroup i is written in terms of the model parameters as follows

$$
\begin{aligned}
P(\mathbf{x}|i) &= \sum_{\mathbf{y}} P(\mathbf{x},\mathbf{y}|i) \\
&= \Phi_{it}\ \Phi_{ix_1} \cdots \Phi_{ix_k} \ .
\end{aligned}
$$

This expression does not contain the attraction parameters. Furthermore, if **x**|j is the set of all latent responses **x** with $X_j=x_j$ and if **y**|j is the set of all observed responses **y** with $Y_j=y_j$, then the attraction parameters for item j in subgroup i are equal to

$$\Phi_{ix_jy_j} = P(y_j|x_j,i) = \{ \sum_{\mathbf{y}|j} \sum_{\mathbf{x}|j} P(\mathbf{x},\mathbf{y}|i)\}/\{ \sum_{\mathbf{y}} \sum_{\mathbf{x}|j} P(\mathbf{x},\mathbf{y}|i)\}.$$

Now the expression does not contain the Rasch model parameters.

In the M-step, the attraction parameters can be directly computed from Equations 3.11 and 3.13 by

$$\Phi_{ix_jy_j} = f_{ix_jy_j} / f_{ix_j} \ , \qquad (j=1,\ldots,k). \tag{3.14}$$

and the Rasch model parameters are computed from the estimated observed counts $f_{i\mathbf{x}}=\Sigma_{\mathbf{y}} f_{i\mathbf{x}\mathbf{y}}$ as a solution to the likelihood equations

$$
\begin{aligned}
m_{ix_j} &= f_{ix_j} \qquad (j=1,\ldots,k), \\
m_{it} &= f_{it}
\end{aligned}
$$

contained in Equations 3.9 and 3.10. This can be done in the usual manner by means of the iterative proportion fitting procedure (Goodman, 1974a, 1974b; Haberman, 1979; Kelderman, 1984). After the parameters of the SERE model are estimated new values for the estimated observed frequencies $f_{i\mathbf{x}\mathbf{y}}$ are computed in the E-step (3.10), and again in the M-step the model parameters will be computed. This procedure is continued until the estimates converge.

3.5.2 Restricted parameters

In the above algorithm the assumption was that the attraction parameters and the item difficulties are different for various groups of subjects. Westers and Kelderman (1992) assumed that items may show DIF both in the attraction parameters and in the latent response, where an item shows DIF in the latent response if equally able subjects from various subgroups have different probabilities of "Knowing" the answer, and an item shows DIF in the attraction parameters if the attractiveness of the alternatives varies from subgroup to subgroup, conditional on their ability. If we assume that an item, say item 1, shows no DIF in the attraction parameters, then for j=1 equation 3.14 is replaced by the equation (Clogg & Goodman, 1985)

$$\Phi_{ix_1y_1} = f_{x_1y_1} / f_{x_1} .$$

Furthermore, if it can be assumed that an item, again say item 1, shows no DIF in the latent response, then across groups the expected frequencies m_{ix} are equated to the corresponding marginal frequencies.

3.5.3 Pseudo-likelihood theory

As already noted, the traditional estimation method can not be used in practice when the number of items is small. The division of the entire item set into several subsets, could lead to a solution of this problem. This means, however, that the parameters of an item in one subset ought to be equal to the parameters of the same item in the other subsets. Therefore, the parameters of the collapsed SERE models have to be estimated simultaneously. By using pseudo-likelihood estimators (Arnold & Strauss, 1988) this requirement can be met.

In this section, the theory of pseudo-likelihood estimation as described in Arnold and Strauss will be discussed and applied to the SERE model. Whenever possible the same notation will be used as in the previous sections.

Let, following Arnold and Strauss, the k-dimensional vector Y_{iv} denote the observed response pattern of the *vth* subject in group i ($i=1,...,g$; $v=1,...,n_i$) on the k items, and let λ denote a coordinate of the p-dimensional parameter space Λ. For the SERE model, the coordinates of Λ are the ability parameters, the difficulty parameters and the attraction parameters for all g groups and all k items. Furthermore, let S denote the class of selected subsets of items and let y_{siv} be the random vector with the coordinates y_{ijv} of y_{iv} for which item j is in the subset s. Finally, denote, still following Arnold and Strauss, the joint density of y_{iv} by $P(y;\lambda,i)$ and the joint density of y_{siv} by $P_s(y_s;\lambda,i)$. For the SERE model, both densities are given by Equation 3.3, but for the second density k is equal to the number of items in the subset s.

According to Arnold and Strauss the pseudo-loglikelihood of the data is then defined as the sum over the subset s of the sum over all observations of the logarithm of the joint densities $P_s(y_s;\lambda)$. A pseudo-likelihood estimator of λ will be a statistic maximizing the pseudo-loglikelihood with respect to λ. This means that if for each subset s, L_s denotes the loglikelihood of the collapsed SERE model, the pseudo-loglikelihood (PL) of the entire SERE model is

$$PL = \sum_s L_s .$$

A pseudo-likelihood estimate of the parameters of the SERE model will be a point in the parameter space for which PL is maximal.

To ensure a solution to the pseudo-loglikelihood, Arnold and Strauss assume that the regularity conditions as mentioned in Theorem 1.1 of Lehmann (1983, p. 406) are met. If these regularity conditions hold, then the solution of the pseudo-loglikelihood equation can be obtained by differentiation of the pseudo-loglikelihood to each element of Λ and the setting of the derivatives to zero. Let for each subset s, f_{sixy} and m_{sixy} be the estimated observed and expected numbers of subjects in subgroup i with latent response vector $\mathbf{x}_s$ and observed response vector $\mathbf{y}_s$ under the collapsed SERE model. Furthermore, if we define

$$m^{(+)}_{ix_jy_j} \equiv \sum_s m_{six_jy_j} ,$$

$$f^{(+)}_{ix_jy_j} \equiv \sum_s f_{six_jy_j} ,$$

for each collapsed SERE model, the pseudo-likelihood equations for the entire SERE model can be formulated as

$$m^{(+)}_{ix_jy_j} = f^{(+)}_{ix_jy_j} , \qquad (j=1,...,k), \tag{3.15}$$

$$m_{sit} = f_{sit} , \tag{3.16}$$

in which m_{sit} and f_{sit} are the expected and estimated observed number of subjects in subgroup i with sum-score t for subset s. Both equations are obtained from the likelihood equations 3.11 and 3.12 for each collapsed SERE model in which interchangability of the operations differentiation and summation is used.

Under the regularity conditions of Theorem 1.1 of Lehmann (1983, p. 406) and assuming that (1) the densities $P_s(y_s;\lambda)$ for different values of λ are essentially distinct, (2) the supports of the densities do not depend on λ and (3) the parameter space Λ contains an open interval ω of which the true parameter λ_0 is the interior point, Arnold and Strauss have shown, with Theorem 2.1, Theorem 2.2 and the arguments in Section 6.4 of Lehmann (1983, p 409-436), that under their regularity conditions a solution of the pseudo-likelihood equations is consistent and asymptotically multivariate normal. This result quarantees that the solutions of the pseudo-likelihoods equations for the SERE model are consistent and asymptotic normal under these conditions.

Finally, as Arnold and Strauss have indicated, pseudo-likelihood estimators are not efficient, but the loss of efficiency may not be large. For the SERE model the lack of efficiency is obvious, since by dividing the set of items into subsets, information about the dependency between items from different subsets is neglected. However, with an optimal choice of the subsets the loss of efficiency may be minimized.

In the next section we will describe how the solutions of the pseudo-likelihood equations 3.15 and 3.16 can be obtained.

3.5.4 The simultaneous estimation method

As discussed in the previous section the maximum pseudo-likelihood estimates can be obtained by solving the Equations 3.15 and 3.16. This can be done in the following way.

First, initial values for the parameters on the right-hand side of Equation 3.6 are determined. How the initial values can be determined will be discussed in Section 3.5.4. Furthermore, let n_{si} be the number of subjects in group i who respond to all the items in subset s and $P_s(\mathbf{x},\mathbf{y}|i)$ be the conditional distribution of latent response $\mathbf{x}_s$ and observed response $\mathbf{y}_s$ for subset s given subgroup i

$$P_s(\mathbf{x},\mathbf{y}|i) = \Phi_{it}\,\Phi_{ix_1} \ldots \Phi_{ix_k}\;\Phi_{ix_1y_1} \ldots \Phi_{ix_ky_k}\,,$$

in which k is equal to the number of items in subset s. In view of the initial values for the parameters, m_{sixy} is computed for each subset s by

$$m_{sixy} = n_{si}\,P_s(\mathbf{x},\mathbf{y}|i)\,, \tag{3.17}$$

Then for each subset s the estimated observed frequencies f_{sixy} are computed by means of

$$f_{sixy} = m_{sixy}\,n_{siy}\,/\,m_{siy}. \tag{3.18}$$

This is the Expectation step (E-step) of the EM-algorithm. Next, in the Maximizaton step (M-step), the model parameters are obtained by solving the Equations 3.15 and 3.16 for the model parameters.

Similar to the case of the traditional estimation method, it can be shown that the attraction parameters and the difficulty parameters can be computed separately. This means that in the M-step the attraction parameters can be directly computed by

$$\Phi_{ix_jy_j} = f^{(+)}_{ix_jy_j} / f^{(+)}_{ix_j}, \qquad (j=1,\dots,k), \tag{3.19}$$

and the Rasch parameters are computed from the estimated observed counts $f_{six} = \Sigma_y f_{sixy}$ by solving the likelihood equations

$$m^{(+)}_{ix_j} = f^{(+)}_{ix_j}, \qquad (j=1,\dots,k), \tag{3.20}$$

$$m_{sit} = f_{sit}, \tag{3.21}$$

contained in the Equations 3.15 and 3.16. Just as in the traditional estimation method for each collapsed SERE model, the Rasch parameters could be computed by means of the IPF procedure; the Rasch parameters, however, are restricted over subsets. For instance, the difficulty of item j in one subset, ought to be equal to the difficulty of the same item in another subset. The Rasch parameters are therefore computed by the following procedure.

Let $\mathbf{x}_s|t$ be the set of all response patterns $\mathbf{x}_s$ in subset s with sum-score t, and $\mathbf{x}_s|x_j$ be the set of all response patterns $\mathbf{x}_s$ in subset s with $X_j=x_j$. In view of the pseudo-likelihood equations 3.20 and 3.21, the item parameters can be derived when the expected marginal counts are written in terms of the parameters and they are equated with the observed counts

$$f^{(+)}_{ix_j} = \sum_s n_{si} \sum_{\mathbf{x}_s|x_j} \Phi_{sit} \prod_{u\in s} \Phi_{ix_u}, \tag{3.22}$$

$$f_{sit} = n_{si} \sum_{\mathbf{x}_s|t} \Phi_{sit} \prod_{u\in s} \Phi_{ix_u}, \tag{3.23}$$

in which Φ_{sit} are the sum-score parameters for sum-score t of group i in subset s. Since the difficulty parameters of item j do not depend on $x_s|x_j$, they can be brought before the summation sign and solved as

$$\Phi_{ix_j} = f^{(+)}_{ix_j} / \sum_{s} n_{si} \sum_{x_s|x_j} \Phi_{sit} \prod_{\substack{u \in s \\ u \neq j}} \Phi_{ix_u} , \tag{3.24}$$

which gives the recursion formula

$$\begin{aligned} \Phi^{(r+1)}_{ix_j} &= f^{(+)}_{ix_j} / \sum_{s} n_{si} \sum_{x_s|x_j} \Phi^{(r)}_{sit} \prod_{\substack{u \in s \\ u \neq j}} \Phi^{(r)}_{ix_u} , \\ &= \Phi^{(r)}_{ix_j} \, f^{(+)}_{ix_j} / m^{(+)(r)}_{ix_j} \end{aligned} \tag{3.25}$$

in which r denotes the iteration number. In a similar way the recursion formula for the sum-score parameters is derived as

$$\Phi^{(r+1)}_{sit} = \Phi^{(r)}_{sit} \, f_{sit} \, / \, m^{(r)}_{sit} \tag{3.26}$$

New estimates of m_{six} can then be obtained by

$$m_{six} = n_{si} \, \Phi_{sit} \prod_{j} \Phi_{ix_j} \tag{3.27}$$

This inner iteration process will be continued until convergence has been reached. After the parameters of the SERE model are estimated, for each subset s, new values for the estimated observed frequencies f_{sixy} are computed in the E-step, and then in the M-step the model parameters will be computed. This procedure is continued until the estimates converge.

If we assume that an item shows no DIF in the latent response or DIF in the attraction parameters, then the algorithm can be adjusted, similar to the one described in Section 3.5.2.

As mentioned before, for the first iteration of the estimation method initial values for the parameters of the SERE model have to be chosen. In the next section this matter will be discussed briefly.

3.5.5 Initial values

Since each collapsed SERE model can be regarded as a latent class model, we can use the Anderson-Lazarsfeld-Dudman method (Anderson, 1954; Lazarsfeld & Dudman, 1951) to obtain initial values for the latent class probabilities (i.e. $P(\mathbf{x}_s)$) and the attraction parameters (Goodman, 1974b). However, to compute initial values for the difficulty parameters and the sum-score parameters from the initial values for the latent class probabilities, a second method is needed. In the literature about the EM-algorithm some suggestions are given for initial values in a number of specific situations (Dempster, Laird & Rubin, 1977; Little & Rubin, 1991).

Below we will describe an alternative method for the above two-stage procedure, but first we will discuss why good initial values may be important.

One reason why good initial values may be important is the rate of convergence of the EM-algorithm. There are several important properties of the EM-algorithm (Dempster, Laird & Rubin, 1977; Rubin, 1991; Wu, 1983), where the one that is most important for this section is the property that the rate of convergence of the EM-algorithm may be painfully slow. In order to alleviate the problem of a slow convergence initial values for the parameters may be chosen that are close to the true values.

Another reason why good initial values may be important is the problem of degenerated solutions (Bartholomew, 1987). Degenerated solutions may occur when the EM-algorithm converges to a solution which lies in the parameter space, but it is not the maximum likelihood solution. An example of a situation where degenerated solutions may occur is the situation where some of the item parameters diverge to infinity. In practice this would mean that the conditional probability of a correct response to an item, given that the subject is in the "Know" state, is indefinite. Such degenerated solutions can be avoided when we try to choose initial values that are close to the true values.

Below we will describe how initial values for the parameters can be determined for the case where it is assumed that a subject will always choose the right alternative (i.e. $Y_j=1$) if (s)he is in the "Know" state. Secondly, we assume that the attractiveness of the correct alternative of item j is equal to $1/r_j$. Thirdly, from Equation 3.1 it follows that the difficulty of an item is equal to the logit of the probability of being in the "Know" state corrected by a constant c. Since this constant c is equal for all items, it can be determined by setting one of the item difficulties equal to zero. In the remainder of this section, the method is only discussed for the pair of items (1,2) (i.e. s=(1,2)). It may be clear that for other pairs of items the method would be similar.

With respect of the above-mentioned assumptions, the determination of the initial values for the parameters can be regarded as a four-step process. First, assuming that the attractiveness of the correct alternative is equal to $1/r_j$, the initial value for the proportion of subjects who were in the "Don't know" state, but answered the item correctly, is set equal to the mean of the proportions of incorrect answers. Given this initial value and the proportions of incorrect answers, the attraction parameters can then be easily computed. For example, the initial value for the attraction parameter of a distractor is equal to the quotient of the proportion of subjects who have chosen that distractor and the initial proportion of subjects who were in the "Don't know" state.

Secondly, the initial proportion of subjects who were in the "Know" state is then equal to the difference between the initial proportion of subjects who are in the "Don't know" state but answered the item correctly and the observed proportion of correct answers. When we multiply these new initial values with the number of subjects n_{si}, we get initial values for the estimated observed number of subjects in group i who were in the "Know" state. We will denote these counts by K_{sij} ($j=1,2$), where j indicates the number of an item in the pair of items. Thirdly, the initial value for the difficulty of an item is then set equal to the logit of the initial proportion of subjects who were in the "Know" state.

For the fourth step, we will assume that the likelihood equation (3.16) holds, which means that the initial values for the sum-score parameters are restricted by this equation and the chosen initial values for the attraction parameters and difficulty parameters. When we write the left-hand side of Equation 3.16 in terms of the model parameters, the initial values for the sum-score parameters can be obtained by

$$\Phi_{sit} = f_{sit} \, / \, n_{si} \, \gamma_t \qquad (3.28)$$

where γ_t is the symmetric function of order t. Given the initial values for the difficulty parameters, only the counts f_{sit} are unknown in this equation. However, these counts can be obtained from Equation 3.23. If the terms of Equation 3.23 are rearranged and $y=(y_1,y_2)$ and $x=(x_1,x_2)$ are equated with (1,1) and (0,0), the marginal counts f_{sit} are the solutions of the following linear system of equations

$$\begin{aligned} f_{si1} + 2\, f_{si2} &= K_{si1} + K_{si2} \\ f_{si0} + \quad f_{si1} + \quad f_{si2} &= n_{si} \qquad (3.29) \\ A\, f_{si0} + B\, f_{si1} + \quad f_{si2} &= n_{siy} \end{aligned}$$

with

$$A = \Phi^{X_1 Y_1}_{x_1 y_1} \, \Phi^{X_2 Y_2}_{x_2 y_2}$$

and

$$B \equiv [\Phi_1^{X_1}\,\Phi_{x_2y_2}^{X_2Y_2} + \Phi_1^{X_2}\,\Phi_{x_1y_1}^{X_1Y_1}]/(\Phi_1^{X_1} + \Phi_1^{X_2})$$

To illustrate the use of the proposed method a data set conforming the SERE model was generated for 17 four-choice items. The item difficulty parameters were chosen from the interval [-2,2]. Latent traits values of 10,000 subjects were drawn from a normal distribution with mean zero and variance one. The attraction parameters of the alternatives for the first nine items were equal to 0.1, 0.2, 0.3 and 0.4 respectively. The attraction parameters of the alternatives for the remaining eight items were equal to 0.4, 0.3, 0.2 and 0.1 respectively. It was assumed that the first alternative (denoted by A) was the correct alternative and that a subject in the "Know" state would always choose the correct alternative.

Table 3.2
Initial values of the item difficulties and the attraction parameters for the alternatives of homogeneous SERE items

	Attraction parameters				Item difficulties	
Item	A	B	C	D	Initial	True
1	.242	.173	.249	.336	4.8365	2.0
2	.250	.171	.251	.328	2.4681	1.5
3	.250	.165	.245	.340	1.2274	1.0
4	.250	.164	.248	.338	0.4903	0.5
5	.250	.161	.248	.341	0.0000	0.0
6	.250	.160	.245	.345	-0.5470	-0.5
7	.250	.162	.250	.338	-1.0051	-1.0
8	.250	.157	.246	.346	-1.4732	-1.5
9	.250	.159	.248	.344	-1.8879	-2.0
10	.250	.384	.246	.120	-1.9741	-1.5
11	.250	.364	.254	.132	-1.5906	-1.0
12	.250	.381	.244	.125	-1.2143	-0.5
13	.250	.386	.235	.129	-0.8529	0.0
14	.250	.373	.252	.125	-0.5556	0.5
15	.250	.374	.258	.118	-0.2021	1.0
16	.250	.370	.254	.126	0.0526	1.5
17	.250	.375	.249	.127	0.2988	2.0

In Table 3.2 the initial values and the real values of the item difficulties and attraction parameters of the SERE model are given. As can be seen from the table for cases with medium to low true item difficulties (i.e. range [-2,.5]) and low attractiveness of the correct alternative (i.e. 0.1) the initial values for the parameters of the SERE model were relatively well determined. For all other cases with low attractiveness of the correct alternative, the initial values for the difficulty of the item were too high. On the other hand, if the attractiveness of the correct alternative was 0.4, for all cases the initial values for the item difficulty were too low. The reason for these discrepancies is that the attractiveness of the correct alternative is assumed to be equal to 0.25. For Item 1 we have modified the algorithm, otherwise the initial value of the number of subjects who were in the "Know" state could be smaller than zero. If the proportion of correct responses is smaller than 0.25 (i.e. $1/r_j$), the initial proportion of subjects who were in the "Know" state was set equal to 0.5%.

In the above derivation of the initial values, several assumptions were made. These assumptions may be wrong. For example, in a similar way as (3.29) was derived, a system of equations can be derived for subsets of three items. This system, however, consists of four equations from which one is nonlinear. For the future, robustness analysis may give answers to the questions whether the proposed method also provides good initial values of the parameters for other cases or whether slightly different initial values still give the same solutions.

3.5.6 Testing the SERE model

Generally, the overall goodness-of-fit of an incomplete latent-class model can be tested by the Pearson statistic (Q) or the likelihood-ratio test statistic (LR) (see Haberman, 1979). Both statistics are asymptotically distributed as chi-square with degrees of freedom equal to the difference between the number of cells in the observed contingency table and the number of parameters estimated. For the selection of the best fitting model one can use the fact that the difference between two likelihood-ratio test statistics of two nested models is also chi-square distributed with the degrees of freedom equal to the difference in the degrees of freedom of the two nested models (Bishop, Fienberg & Holland, 1975).

An alternate approach to model selection, also based on the likelihood principle, was developed by Akaike (1977, 1987). Akaike's Information Criterion (AIC) for a model with likelihood L, is defined as $AIC = -2\ln(L)+2D$, where D is the number of independent parameters estimated in fitting the model. The first term of AIC is a measure of badness of fit, whereas the second term is a penalty term correcting for overfitting due to the increasing bias in the first term as the number of parameters in the model increases. The model with the minimum AIC value is chosen as the best fitting model.

As can be seen from the definition, AIC is inconsistent in the sense that an increase of the sample size does not have a direct impact on the criterion. To reflect the sample size in the penalty term Bozdogan (1987), Raftery (1986a, 1986b) and Schwarz (1978) presented some other information criteria. The consistent version of the AIC of Bozdogan has ln(n)+1 (where n is the sample size) as a multiplication factor for the number of independent parameters, whereas the multiplication factor in the information criteria of Raftery and Schwarz is equal to ln(n). For large sample sizes, the consistent criteria have a larger penalty term than AIC. Consequently, the consistent criteria tend to lead to simpler models than AIC does.

For pseudo-likelihoods, information criteria can be defined based on the same notion as on with AIC is based, namely the minimization of the Kullback-Leibler (1951) information quantity. Let $\mathbf{Y}_s$ be a continuous random vector characterized by a known probability density $P_s(y_s;\lambda)$. In the case of the SERE model, $\mathbf{Y}_s$ denotes the observed response pattern on the k_s ($k_s<k$) items in the subset s, the elements of λ are the sum-score parameters, the difficulty parameters and the attraction parameters for all k items, and the density $P_s(y_s;\lambda)$ is given by Equation 3.3, but with k equal to the number of items (k_s) in the subset s. Furthermore, let us assume that there is a true parameter vector λ^*. Finally, let us suppose that all the competing models are generated by simply restricting the parameter vector λ.

The objective of the estimation procedure and the model selection procedure is then to select λ closest to the true parameter vector λ^*. We will measure the closeness or the goodness-of-fit by means of Kullback-Leibler information quantities

$$I(\lambda^*;\lambda) \equiv \sum_{s} E_{s*}[L_{s*}-L_s],$$

where, for each subset s, L_s and L_{s*} are the loglikelihoods of the estimated and the true parameters, and E_{s*} denotes the expectation with respect to the true distribution $P_s(y_s;\lambda^*)$. If we denote D_s as the number of independent parameters in the collapsed SERE model, then for every subset s, $AIC(s) = -2L_s+2D_s$ is a natural estimator of the following quantity $-2E_{s*}[L_s]$ (Bozdogan, 1987). Minimizing $I(\lambda^*;\lambda)$ would then be equal to searching for a model that minimizes the sum over all subsets s of $AIC(s) + E_{s*}[L_{s*}]$. Since $E_{s*}[L_{s*}]$ is a constant term for all competing models, searching for the best fitted model will be equal to searching for a model that minimizes the sum over all subsets s of AIC(s).

Since we have used the pseudo-likelihood theory for the estimation of the parameters of the SERE model, the total number of degrees of freedom (D) is not equal to the sum of the degrees of freedom (D_s) of all collapsed SERE models, but equal to the number of independent parameters in the entire SERE model. For all collapsed SERE models, $[\Sigma_s k_s]$ sum-score parameters, k-1 difficulty parameters and $\Sigma_j (r_j-1)$ attraction parameters have to be estimated. Therefore, D must be equal to the sum of these numbers.

Further, given the definition that the pseudo-loglikelihood (PL) is equal to the sum over the subsets s of the loglikelihoods (L_s), we will now be able to define a pseudo Akaike information criterion (PAIC) as

$$\text{PAIC} = -2\text{PL} + 2\text{D}$$

as a measure of the relative 'distance' between the true parameter vector λ^* and the model parameter vector λ. In our future analyses, the model with the minimum PAIC will be chosen to be the best fitting model.

We can adjust PAIC to make it consistent by changing the multiplication factor 2 in the penalty term into ln(n)+1 (Bozdogan, 1987), ln(n) (Raftery, 1986a, 1986b; Schwarz, 1978) or any other function depending on n (Sclove, 1987).

Since the derivations of the (pseudo) information criteria are based on likelihood ratio test statistics, objections can be raised to their use since asymptotic results may not hold. Since some analytical conditions, required for the proper use of the (consistent) pseudo information criteria, may not be met, more research is required before these information criteria may be regarded as measures of quality.

3.6 APPLICATIONS OF THE DBI-PRINCIPLE

For the estimation of the parameters of the SERE model when the number of items is large, the computer program LANPACO (Westers & van der Sar, 1993; Appendix A.3) was written. LANPACO is a Turbo Pascal program which calculates not only the estimates of the parameters in the SERE model by using the DBI-principle, but which also has an user-interface which provides graphical display of the results. Furthermore, LANPACO automatically selects all possible pairs of items as subsets of items.

To illustrate the use of the proposed estimation method, two test data sets which conformed with the SERE model and the Rasch model were generated for 17 items. The item difficulties were chosen from the interval [-2,2]. Latent traits values of 10,000 subjects were drawn from a normal distribution with mean zero and variance one. All 17 items, which conform with the SERE model, were four-choice items, where the attraction parameters of the alternatives for each of the first nine items were equal to 0.1, 0.2, 0.3 and 0.4 respectively. The attraction parameters of each of the alternatives for the last eight items were equal to 0.4, 0.3, 0.2 and 0.1 respectively. We assumed that the first alternative (denoted by A) was the correct alternative and that a subject in the "Know" state would always choose the correct alternative. Please note that the data which conform with the SERE model were simulated under the same conditions as the simulated data in Section 3.5.5.

In Table 3.3 the real item difficulties and the estimated item difficulties of all 17 items for both sets of data, as well as the estimated attraction parameters of the set of data which conform with the SERE model, are given. The item difficulty estimates and the attraction parameter estimates were obtained through the LANPACO program, and for both sets of data the item difficulty of the fifth item was equated with its real value of zero. The iteration process was continued until the maximum of the absolute difference between the new and the old values of the parameter estimates was smaller than 0.00001.

Table 3.3

Estimated values of the item difficulties and the attraction parameters for the alternatives of homogeneous SERE and Rasch items

	Attraction parameters SERE data				Item difficulties		
Item	A	B	C	D	SERE	Rasch	True
1	.030	.221	.319	.431	1.3259	1.9658	2.0
2	.035	.220	.323	.422	1.1106	1.5421	1.5
3	.046	.209	.311	.433	0.7373	1.0050	1.0
4	.067	.203	.309	.421	0.3384	0.5120	0.5
5	.091	.195	.300	.414	0.0000	0.0000	0.0
6	.134	.184	.283	.398	-0.4243	-0.5033	-0.5
7	.198	.172	.267	.362	-0.7796	-0.9597	-1.0
8	.288	.148	.234	.330	-1.1121	-1.5019	-1.5
9	.379	.130	.205	.286	-1.4170	-1.9786	-2.0
10	.397	.310	.198	.095	-1.4454	-1.5022	-1.5
11	.319	.332	.231	.119	-1.1745	-0.9506	-1.0
12	.255	.380	.243	.123	-0.8669	-0.4701	-0.5
13	.195	.416	.252	.138	-0.6060	0.0175	0.0
14	.156	.421	.283	.140	-0.3860	0.5126	0.5
15	.126	.437	.301	.137	-0.1143	1.0046	1.0
16	.110	.440	.302	.149	0.0679	1.5285	1.5
17	.094	.453	.301	.152	0.2227	2.0543	2.0

As can be seen from Table 3.3, the range of the difficulties of the 17 items which conform with the SERE model decreased from [-2;2] to [-1.4454;1.3259]. Furthermore, for the cases with medium to high true item difficulties [i.e. range [-1,2]) and low true attractiveness of the correct alternative (i.e. 0.1) the parameters of the SERE model were estimated relatively well. Moreover, for the cases in which the true item difficulties were low (i.e. ≤ -1) and the true attractiveness of the correct alternative was high (i.e. 0.4) the parameters were estimated relatively well too. In all other situations the parameters of the SERE were badly estimated. For instance, for medium or large values of the item difficulty and a high attractiveness of the correct alternative, the item was estimated to be easier than was simulated, and the attraction parameter was estimated to be smaller than was simulated. For low values of the item difficulty and a low attractiveness of the correct alternatives, the item difficulty was underestimated and the attraction parameter of the correct alternative was overestimated. These results indicate that a trade-off may have existed between the item difficulties and the attraction parameters.

If we take a closer look at the SERE model, we can see that the specified model is a special case of a three-parameter logistic model. It is assumed that a subject will always choose the correct alternative if the subject is in the "Know" state. Therefore, for the estimation of the item difficulties and the sum-score parameters the observed responses variable Y_j may be dichotomized into a new response variable Z_j, with $Z_j = 1$ if $Y_j = A$ (i.e. the correct alternative) and $Z_j = 0$ for all other observed responses Y_j. The probability of a correct response is then

$$P(Z_j = 1) = \Phi_{0\ A}^{X_jY_j} + (1 - \Phi_{0\ A}^{X_jY_j})\, P(X_j = 1)$$

and this equation is nothing else but a special case of the three-parameter logistic model, in which the discrimination parameter is being held equal to 1.

Literature about the three-parameter logistic model (Baker, 1987; Hambleton & Swaminathan, 1985; Lord, 1980) shows that the properties of the item parameter estimators for the one- or two-parameter logistic models are generally better than those for a three-parameter logistic model. For instance, the three-parameter logistic model does not have sufficient statistics for estimating the parameters. Moreover, for obtaining reliable estimates of the guessing parameter (i.e. the attraction parameter of the correct alternative) many subjects at a low ability level will be required. Finally, Thissen and Wainer (1982) state that "the use of an unrestricted maximum likelihood estimation for the three parameter model either yields results too inexact to be of any practical use, or requires samples of such enormous size so as to make them prohibitively expensive" (p. 403).

In view of the phenomenon of biased parameter estimates for the three-parameter logistic model, it may be expected that the parameter estimates for the specified SERE model are also biased. However, for certain combinations of the item difficulties and attractiveness of the correct alternative the parameter estimates may be less biased (e.g. Item 10).

If the guessing parameter (i.e. the attractiveness of the correct alternative) is set to zero, the SERE model can be viewed as a Rasch model. An example of this kind of data are the 17 items which conformed with the Rasch model. As can be seen from Table 3.3 all item difficulties are very well estimated.

The use of the proposed estimation method needs further study. For instance, with a simulation study it should be examined if the estimates through the proposed estimation method differ not only from those through the traditional estimation methods but also differ from the true parameter values. It should also be examined under which conditions these deviations may be negligible.

3.7 DISCUSSION

In this chapter a new estimation method for the solution-error response-error (SERE) model for a large set of items was proposed. The main idea of the new method is that the entire item set is divided into several subsets. It was shown that the SERE model can then be rewritten into a related set of smaller SERE models. When pseudo-likelihood theory is used, estimates of the parameters of the entire SERE model can then be found. A pseudo-loglikelihood can be expressed as the sum of the loglikelihoods for the smaller models over the subsets. The estimates of the parameters of the SERE model can then be found by maximizing the sum of the loglikelihoods of the smaller SERE models. The main advantages of this approach are the decreased number of latent classes, the decreased numbers of cells in the observed and expected contingency table, and a more efficient use of the data. A disadvantage is that information about the joint relationships among the items may be lost when the SERE model is collapsed.

An important issue with respect to the pseudo-likelihoods concerns the goodness of fit of the SERE model. The likelihood-ratio test statistics for each collapsed SERE model is chi-square distributed with degrees of freedom equal to the difference between the number of cells of the observed contingency table and the number of estimated parameters of the collapsed SERE model. However, an important question is if the (weighted) sum of these likelihoods-ratio test statistics over all subsets is chi-square distributed, or if we can develop other test statistics for the SERE model, like the Martin-Löf (1973) statistic, the statistics of van den Wollenberg (1972, 1982), or the statistics of Glas (1989). Future research should address this question too.

In this chapter, pseudo information criteria were introduced, which were based on the same notion as the one on which the Akaike's information criterion is based. However, future research should address the question whether these pseudo information criteria are of any practical use for the selection of the best fitting model.

Finally, the objective of this chapter was the development of an estimation method that computes SERE models with a large item set much more efficiently, in terms of both computer storage and processing time. In the previous section it was demonstrated that with the proposed estimation method it is possible to estimate the parameters of a SERE model with an item set of 17 four-choice items. At this stage, however, the computer program LANPACO, in which the proposed estimation method was implemented, can handle any number of items as long as the total number of subsets does not exceed 255. Since LANPACO selects all possible pairs of items as subsets of items, this means that the maximum number of items LANPACO can handle lies between eight items for SERE models with eight subgroups and 23 items for SERE models with one subgroup. As indicated by Westers and Kelderman (1992) the traditional estimation method, as implemented in LCAG, can handle only a maximum of four items for SERE models with two subgroups.

Since the LCAG version which was been used is a program that runs VAX system running under VMS and LANPACO is a program that runs under MS-DOS, it is difficult to compare the traditional estimation method and the proposed estimation method with respect to the processing time (i.e. CPU time). However, if we compare the number of multiplications and summations required for estimating the parameters in both estimation methods, some subjective statements about the processing times can be made.

In the following example a test was subjected to one group of examinees (i.e. $g=1$). This test consisted of k items in which each item has r response alternatives (i.e. $r_j=r$ for all $j=1,...,k$). Furthermore, let h be the number of selected pairs of items. In the case of LANPACO, h is equal to $k(k-1)/2$. In view of these choices, for both estimation methods the number of multiplications and summations needed for the computation of the parameters of the SERE model can be approximated. In Table 3.4 the number of multiplications and summations in the computation of the attraction parameters are given for each iteration cycle of both estimation methods.

Table 3.4

Number of multiplications and summations required for each iteration cycle of the traditional estimation method and the proposed estimation method to calculate the attraction parameters of the SERE Model

Traditional estimation method			Proposed estimation method		
Equation	Multiplications	Summations	Equation	Multiplications	Summations
3.13	$(2r)^k(2k+1)$	-	3.17	$20hr^2$	-
3.9	$2(2r)^k$	-	3.18	$8hr^2$	-
3.14	$2rk$	$k(2r)^k+2rk$	3.19	$2rk$	$2rk(2rh+1)$

Analogously, for both estimation methods the number of multiplications and summations needed for the computation of the item difficulties can be obtained. When we compare these numbers, it seems that the proposed estimation method requires a smaller number of multiplications and summations for the estimation of the attraction parameters and item difficulties than the traditional estimation method does. This means that we may expect that the processing time of the proposed estimation method is shorter than the processing time of the traditional estimation method. Experience obtained by the application of the DBI principle as discussed in Section 3.5.5 and obtained during the simulation study of Chapter 4, indicates that the processing time of the proposed estimation method is about 1 to 10 minutes, dependent on the number of items, the number of alternatives, the criterion on which the iteration process will be stopped and, of course, the data. However, experience obtained during the analyses of Westers and Kelderman (1992) indicates that the processing time of the traditional estimation method varies from 10 to 150 minutes, dependent on the restrictions of the postulated SERE model.

In order to summarize, the estimation method based on the pseudo-likelihood theory provides not only consistent and asymptotic normal estimators of the parameters, but it is also much more efficient, in terms of both computer storage and processing time, than the traditional estimation method. The only drawback is that the estimators cannot be expected to be asymptotically efficient.

Chapter 4

A SIMULATION STUDY OF THE SOLUTION-ERROR RESPONSE-ERROR MODEL

4.1 INTRODUCTION

In this chapter the results from a simulation study of the solution-error response-error (SERE) model of Kelderman (1988, see also Westers & Kelderman, 1992) and of the estimation technique presented in Chapter 3 are reported. The questions considered are: (1) Can differential item functioning (DIF) still be found if the number of items or the number of subjects is small?; (2) How do the values of the estimators differ from the true parameters?; (3) Is this deviation consistent in the sense that the differences tend to decrease when the number of subjects increases? With simulation we will also examine under which conditions the SERE model can be used in practice and whether DIF can be detected. However, it must be stressed that this study does not pretend to be a systematic and comprehensive study of the robustness of the estimation method or the quality of the SERE model.

Section 4.2 is devoted to a brief description of the SERE model and the estimation method from Chapter 3. In Section 4.3 the research questions of the simulation study are discussed, whereas in Section 4.4 a complete description of the simulated data is given. Finally, in Section 4.5 the results of the simulation study will be discussed.

4.2 THE SOLUTION-ERROR RESPONSE-ERROR MODEL

In the solution-error response-error (SERE) model (Kelderman, 1988), a distinction is made between two states: a "Know" state and a "Don't know" state. The states determine whether the subject can or cannot solve the problem imposed by an item. The probability that the subject is in the "Know" state is assumed to be governed by the Rasch (1960/1980) model. Furthermore, the assumption is that if the subject is in the "Don't know" state, (s)he will choose the most attractive alternative, where the attractiveness of an alternative may be dissimilar for different alternatives, including the correct one.

Let x_{nj} ($x_{nj} = 0,1$) and y_{nj} ($y_{nj} = 1,\ldots,r_j$) indicate the latent response and observed response of subject n (n = 1,...,N) to item j (j = 1,...,k), respectively. The random variables associated with x_{nj} and y_{nj} are denoted by X_{nj} and Y_{nj}, respectively. Assuming that the latent response is governed by the Rasch model, the probability of x_{nj}, given that the subject n has ability θ, is

$$(4.1) \qquad P(x_{nj} \,|\, \theta) = \exp(x_{nj}(\theta-\delta_j))/[1 + \exp(\theta-\delta_j)] .$$

Furthermore, the assumption is that the relationship between the latent response x_{nj} and the observed response y_{nj} is the same for each subject n and described by the conditional probability

$$(4.2) \qquad \Phi^{X_jY_j}_{x_jy_j} \equiv P(y_{nj} \,|\, x_{nj}) ,$$

in which the superscripts, in symbolic notation, indicate that the random variables X_j and Y_j are involved in the conditional probability. This conditional probability will be referred to as the attraction parameter of item j.

Finally, assuming that y_{nj} only depends on x_{nj} and that x_{nj} only depends on the latent ability θ, we have

$$(4.3) \qquad P(y_{nj} \,|\, \theta) = [\Phi^{X_jY_j}_{0\,y_j} + \Phi^{X_jY_j}_{1\,y_j} \exp(\theta-\delta_j)]/[1+\exp(\theta-\delta_j)] .$$

One of the main advantages of the SERE model is that it can be easily formulated as a latent class analysis model (Kelderman, 1988), namely as a latent class model in which the structure of the latent-class probabilities is explained by a loglinear Rasch model. Each latent class corresponds with an idealized response pattern. Another advantage of the SERE model is that, by extending the SERE model with variables defining subgroups, it is not only possible to test whether a certain item shows DIF, but also to test whether this DIF is caused by the difficulty of the item, the attractiveness of the alternatives, or both (Westers & Kelderman, 1992). Generally, an item shows DIF if the probability of a correct response among equally able test takers is different for various racial, ethnic, or gender subgroups. However, an item can show DIF in two different ways. In the first place, an item shows DIF if equally able subjects from different subgroups have different probabilities of "Knowing" the answer to the problem imposed by the item. Secondly, an item also shows DIF if the attractiveness of the alternatives of the item varies from subgroup to subgroup conditioned on ability. Westers and Kelderman (1992) refers to these two types of DIF as DIF in the latent response and DIF in the attraction parameters. They also show that both types can be examined with the SERE model.

As discussed in Kelderman (1988, Westers & Kelderman, 1992), the parameter estimates can be computed with the methods LCAG (Hagenaars, 1988; Hagenaars & Luijkx, 1990) and LOGIMO (Kelderman & Steen, 1988). LCAG is a computer program for the estimation of the parameters of loglinear models with latent variables. Apart from the estimated attraction of the alternatives, it also gives the estimated expected frequency distribution of the latent classes under the SERE model. LOGIMO is a general computer program for analyzing loglinear IRT models. We use it here to compute the difficulty of the items from the frequency distribution of the latent classes in the SERE model.

The use of these two methods, however, becomes unpractical for a large number of polytomous items. In the first place, in LCAG all cell frequencies, including empty cells with frequency zero, have to be stored. Secondly, in LCAG the values for the probability of the latent classes followed by the values for the conditional probabilities of the observed variables given each latent class, have to be stored. Doing this for the case of the SERE model, LCAG uses such a large amount of memory space that it is impossible to consider a large item set. For example, Westers and Kelderman (1992) could only consider four five-choice items at a time.

Therefore, in Chapter 3 a maximum likelihood estimation method for the SERE model was proposed, which is based on the division of the entire item set into several subsets of items. It was shown that the SERE model can then be rewritten into a set of smaller SERE models. We will refer to these smaller SERE models as the collapsed SERE models. With the use of pseudo-likelihoods, estimates of the parameters of the entire SERE model can be found. A pseudo-loglikelihood could be expressed as the sum of the true loglikelihoods for subsets of items. The estimates of the parameters of the SERE model could then be found through the maximization of the sum of the loglikelihoods of the collapsed SERE models (Chapter 3). The advantages of this approach are the decreased number of latent classes, the decreased number of cells in the observed and expected contingency table, and a more efficient use of the data. More efficient use of the data because, apart from designs with complete data, designs with incomplete data can also be used. Data from any subject, even when responding to only two items of the test, can be studied. Of course the data for these subjects can only be used if the two items form one of the subsets of items. A disadvantage of the approach of Chapter 3 is that some of the statistical efficiency of the estimators may be lost.

The overall goodness of fit of the collapsed SERE model can be tested by the Pearson statistic or the likelihood-ratio test statistic. With the difference of the likelihood-ratio test statistics for two nested models the best fitted model can be selected (Bishop, Fienberg, & Holland, 1975; Rao, 1973). In Chapter 3 an alternate approach to model selection is described which uses the pseudo-likelihood estimates and some modified versions of the information criteria of Akaike (1977, 1987), Bozdogan (1987) or Raftery (1986a, 1986b). With these so-called pseudo Akaike's information criteria, it can be checked whether a model gives a significant improvement in fit over another model.

4.3 RESEARCH QUESTIONS OF THE SIMULATION STUDY

It is well-known that the Pearson statistic and likelihood-ratio test statistic for testing the overall goodness-of-fit of a model are both asymptotically distributed as chi-square with degrees of freedom equal to the difference between the number of cells in the observed contingency table and the number of estimable parameters. However, by using the pseudo-likelihood theory for the estimation of the parameters of the SERE model, the use of the Pearson or likelihood-ratio goodness-of-fit statistics is not allowed.

On the other hand, there are other indices which can be used for the selection of the best fitted model. The information criteria of Akaike (1977, 1987), Bozdogan (1987) or Raftery (1986a, 1986b) for example. Akaike's information criterion (AIC) for a model with likelihood L is defined as $AIC = -2\ln(L)+2D$, in which D is the number of independent parameters which are estimated in fitting the model. The model with the minimum AIC value is chosen to be the best fitting model. Since AIC is inconsistent in the sense that an increasing sample size does not have a direct impact on the criterion, modifications of the criterion are proposed in the literature. For example, the consistent AIC criterion (CAIC) of Bozdogan has $\ln(n+1)$ (i.e. n is the sample size) as a multiplication factor for the number of independent parameters, whereas the multiplication factor in the Raftery's Bayesian information criterion (BIC) equals $\ln(n)$. Generally, the CAIC and BIC criteria tend to lead to simpler models than AIC does. With the use of pseudo-likelihoods, pseudo information criteria can be defined, based on the same notion as those on which the AIC, CAIC and BIC are based, but in which the loglikelihood $\ln(L)$ is replaced by the pseudo-loglikelihood PL and D is equal to the number of independent parameters in the entire SERE model (Chapter 3). The model with the minimum pseudo Akaike information criterion (PAIC) value will be chosen as the best fitting model.

Since the derivations of the (pseudo) information criteria are based on likelihood ratio test statistics, an objection can be raised to their use, because asymptotic results may not be valid. There is, nevertheless, considerable value in studying the behavior of the (consistent) pseudo Akaike information criteria, since their performance in real-life situations may be of practical use. In this simulation study we will therefore examine whether the (consistent) pseudo information criteria of Chapter 3 can be used for the examination of DIF. Since a consistent PAIC tends to lead to simpler models than AIC does, only the following consistent pseudo Akaike's information criterion will be calculated in this chapter

$$\mathrm{PAIC} = -2\ \mathrm{PL} + \ln(n)D$$

If the pseudo-likelihood-ratio test statistic PLR is equal to the sum over all subsets of the likelihood-ratio test statistics of the collapsed SERE models, -2 PL is equal to PLR + C, in which C only depends on the observed data. Since in the simulation study we only compare models with each other for the same data, the PAIC-C values will be reported in the tables.

Another issue in the field of the examination of DIF is the number of items of the test. Generally, DIF is not necessarily some inherently "bad" characteristic of an item; it is also dependent on the pool of items with which the particular item is being compared (Berk, 1982). For instance, biased items can be identified as those that are relatively more difficult for members of a particular group. Since the DIF detection methods all rely on the total test as a measure of the ability, bias will go unnoticed by these methods when all the items have the same type and degree of invalidity. Furthermore, with a small sample of items, it may be difficult to distinguish between systematic differences between groups due to DIF and systematic differences between groups due to ability. Since the biased items contribute to the estimation of the subjects' ability, the DIF detection methods based on IRT models is sensitive for many biased items; too many biased items would ordinarily harm the stability of findings (Shepard, Camilli, & Williams, 1984). However, as Rudner, Getson and Knight (1980b) showed in their paper, the correlations between detected bias and true bias increase only slightly with increasing test length. Two remarks have to be made with respect to this conclusion. In the first place, in their study, data were generated which conform with the three parameter logistic model in which the degree and the type of DIF were specified in advance. Secondly, allmost all considered DIF detection methods, including the method based on the three parameter logistic model, showed a slight general increase in the average correlation with increasing test length. In the simulation study we will examine whether these conclusions are also valid for the SERE model and the pseudo-likelihood estimation method from Chapter 3.

Finally, in the simulation study the attention will be focused on the combination of the two types of DIF: DIF in the latent response and DIF in the attraction parameters. Westers and Kelderman (1992) argued that it is possible to define these two types of DIF by using the SERE model. But is it really possible to detect DIF in the latent response, if DIF already exists in the attraction parameters? Or to detect DIF in the attraction parameters, if the item already shows DIF in the latent response?

The above research questions can be summarized as follows: (1) How do the values of the estimators differ from the true parameters?; (2) Is this deviation consistent in the sense that the differences tend to decrease when the number of subjects increases?; (3) Can DIF still be found if the number of subjects is small?; (4) Can DIF still be found if the number of items is small?; (5) Is it possible to detect an item which shows DIF in the latent response, but shows DIF in the attraction parameters as well?

4.4 THE SIMULATED DATA

The usefulness of the SERE model and the maximum likelihood estimation method from Chapter 3 will be studied with the use of simulated data. It should be noticed in advance that there are several possible combinations of sample sizes, test lengths, number of alternatives, choice of parameter (e.g. difficulty parameter or attraction parameter) values, choice of subsets, choice of items which show DIF (e.g. DIF in the latent response or DIF in the attraction parameter). The simulation study will concentrate only on some interesting combinations of these variables, relevant for each research question. In the next section these combinations will be described in more detail.

In order to generate data which conform with the SERE model, the following algorithm was used.

Input: the sample size N, test length k, number of alternatives vector **r**, the difficulty parameter vector δ, and the attraction parameter matrix Φ.

Step 1: for $n = 1,\ldots,N$, draw θ_n (i.e. the ability of subject n) from the standard normal $N(0,1)$ distribution.

Step 2: for $n = 1,\ldots,N$ and $j = 1,\ldots,k$, draw μ_{nj} from the uniform distribution on $[0,1]$.

Step 3: for $n = 1,\ldots,N$ and $j = 1,\ldots,k$, generate latent responses using **if** $\mu_{nj} < P(X_{nj} = 1 \mid \theta_n, \delta_j)$ **then** $X_{nj} = 1$ **else** $X_{nj} = 0$.

Step 4: for $n = 1,\ldots,N$ and $j = 1,\ldots,k$, draw μ_{nj} from the uniform distribution on $[0,1]$.

Step 5: for $n = 1,\dots,N$, $j = 1,\dots,k$ and $i = 2,\dots,r_j$, generate observed responses

$$\text{if } \sum_{h=1}^{i-1} \Phi^{X_jY_j}_{x_{nj}h} < \mu_{nj} \leq \sum_{h=1}^{i} \Phi^{X_jY_j}_{x_{nj}h} \text{ then } Y_{nj} = i.$$

With this algorithm, different data sets were generated for two groups. By specifying unequal values of the difficulty parameter for both groups it is possible to generate items which show DIF in the latent response. Items with DIF in the attraction parameters can be generated by specifying unequal attraction parameters for both groups.

For convenience during the entire simulation study the same item characteristics will be used. Furthermore, we assumed that if the subject is in the "Know" state, the subject will choose the correct alternative (denoted by A). In Table 4.1 the manifest difficulty parameters and the attractiveness of the alternatives for the "Don't know" state are given. Please note that the number of response categories is taken to be the same for all items, i.e. $r_j = 4$, for $j = 1,\dots,9$. Table 4.1 shows that Items 4 and 8 show DIF in the latent response, whereas Items 2 and 4 show DIF in the attraction parameters.

For the estimation of the parameters of the SERE model the computer program LANPACO was used (Westers & van der Sar, 1993). A description of the program will be given in Appendix C.

Table 4.1
Item parameters of the simulated data

	Group 1					Group 2				
		Attraction parameters					Attraction parameters			
Item	Item difficulty	A	B	C	D	Item difficulty	A	B	C	D
1	2.0	.25	.25	.25	.25	2.0	.25	.25	.25	.25
2	1.5	.25	.25	.25	.25	1.5	.10	.20	.30	.40
3	1.0	.25	.25	.25	.25	1.0	.25	.25	.25	.25
4	0.5	.25	.25	.25	.25	1.0	.10	.20	.30	.40
5	0.0	.25	.25	.25	.25	0.0	.25	.25	.25	.25
6	-0.5	.25	.25	.25	.25	-0.5	.25	.25	.25	.25
7	-1.0	.25	.25	.25	.25	-1.0	.25	.25	.25	.25
8	-1.5	.25	.25	.25	.25	-2.0	.25	.25	.25	.25
9	-2.0	.25	.25	.25	.25	-2.0	.25	.25	.25	.25

4.5 RESULTS

In this section the results of the simulation study will be presented. In Sections 4.5.1 and 4.5.2 the small-sample behavior of the estimation method and the SERE model will be discussed. The issue of the influence of the test length on the examination of DIF will be discussed in Section 4.5.3. Finally, in Sections 4.5.4 and 4.5.5 we will discuss whether it is really possible to examine items that show DIF in the latent response and DIF in the attraction parameters.

4.5.1 Small-sample behavior of pseudo-likelihood estimates

The purpose of the first part of the simulation study was to get some idea of the small-sample behavior of the estimation method and the SERE model. In order to produce an example of the small-sample behavior, the first group and Items 2, 5, 6 and 8 as described in Table 4.1 were chosen. The choice of these four items was based on the following consideration. Generally, a test may have items that show DIF in the latent response (e.g. Item 8), but also items which show DIF in the attraction parameters (e.g. Item 2). Item 5 is chosen as a reference item, because it has a zero difficulty parameter. We set its parameter to zero to fix the scale. Finally, Item 6 is chosen because of its low true difficulty parameter. Sample sizes of 1000, 2000 and 5000 respondents were used and for every sample size 25 replications were made. The estimated values of the parameters of the items are shown in Table 4.2.

Table 4.2
Mean and their standard deviations (SDV) of the attractiveness of the correct alternative and the item difficulties of SERE-homogeneous data for different sample sizes.

		N = 1000		N = 2000		N = 5000	
Item	True	Estimated	SDV	Estimated	SDV	Estimated	SDV
				Attraction parameters			
2	.25	.0649	.0038	.0640	.0032	.0637	.0014
5	.25	.1458	.0093	.1400	.0066	.1356	.0048
6	.25	.1990	.0131	.1916	.0084	.1852	.0073
8	.25	.3652	.0222	.3547	.0168	.3403	.0123
				Item difficulties			
2	1.5	0.776	.0853	0.809	.0592	0.829	.0316
5	0.0	0.000	-	0.000	-	0.000	-
6	-0.5	-0.291	.0631	-0.317	.0587	-0.325	.0300
8	-1.5	-0.844	.0842	-0.916	.0549	-0.943	.0511

In Table 4.2 the true and the mean of the estimated attraction parameters of the right alternatives and the estimated difficulty parameters are given for each sample size. Furthermore, the column labelled "SDV" gives the values of the standard deviation of the estimated attraction parameters and the estimated difficulty parameters, respectively.

Generally, the test score of a subject is determined by the number of correct choices of the right alternative. Therefore, in Table 4.2 only the estimated attraction parameters of alternative A (i.e. the right alternative) will be compared for different sample sizes. For the other alternatives similar tables can be made.

As can be seen from this table the standard deviations of the estimates of the attraction parameters decrease with increasing sample size. The simulation study also shows that except for easy items (e.g. item 8) the difference between the true and the estimated attractiveness of the correct alternative increases with increasing sample size. Table 4.2 shows that the standard deviations of the estimated difficulties, as well as the difference between the true and the estimated difficulties, decreases with increasing sample size.

As discussed in Chapter 3, it was to be expected that the maximum pseudo-likelihood estimates would be less efficient, but consistent. The results of Table 4.2 suggest, however, that the estimates are inconsistent and efficient. High efficiency is not surprising, because LANPACO selects all possible pairs of items as subsets of items, which means that the covariances between the items are not neglected. Neglecting the dependencies between the items would generally decrease the efficiency of the maximum pseudo-likelihood estimates.

If we take a closer look at the results, we will see that for all sample sizes a trade-off exists between the attractiveness of the correct alternative and the difficulties of the items: if the attractiveness of the correct alternative is estimated too low this is compensated by estimating the item difficulties too low, and vice versa. Since the SERE model as defined in this simulation study can be regarded as a special case of a three-parameter logistic model, this trade-off was to be expected (Chapter 3). The literature about the three-parameter logistic model also shows that there is only empirical evidence that consistency of the item parameters may comply with the theoretical expectations (Swaminathan & Gilford, 1983; Wingersky & Lord, 1984). In view of the phenomena of biased estimates for the three-parameter logistic model (Baker, 1987; Hulin, Lissak & Drasgow, 1982; Lord, 1975; Thissen & Wainer, 1982), the item parameter estimates for the SERE model may be expected to be biased as well. And, in particular, the estimation of the difficulties may be affected as an error in the attractiveness of the correct alternative results in a shift in the estimate of the item difficulties. The inconsistency of the results is therefore caused by the structure of the postulated SERE model.

4.5.2 Small-sample behavior of DIF detection

This section deals with the question of the way in which the examination of DIF is influenced by the number of subjects. In order to answer this question, data were generated according to the algorithm of Section 4.2. Just as in the case of the previous part of the simulation study Items 2, 5, 6 and 8 from Table 4.1 were chosen. However, this time the data were generated for two groups. Furthermore, samples sizes of 1000, 2000 and 5000 were chosen and for every sample size 25 replications were made. Finally, the parameters of four models were estimated: (a) a model in which none of the items shows DIF, (b) a model in which Item 8 shows DIF in the latent response, (c) a model in which Item 2 shows DIF in the attraction parameters, and (d) a model in which Item 8 shows DIF in the latent response and Item 2 shows DIF in the attraction parameters. Please note that Model d is the same as to the model under which the data were simulated. In Table 4.3 the values of the consistent pseudo Akaike's information criterion (PAIC) of Model a through d are given for each of three different sample sizes: 1000, 2000, and 5000 subjects, respectively. Furthermore, the columns labelled "SDV" give the values of the standard deviation of these PAIC values, whereas the columns denoted with "Best Model" give the number of analyses in which the particular model has the lowest PAIC value of the four models.

Table 4.3
Mean of the consistent pseudo Akaike's information criteria (PAIC) with their standard deviation (SDV), the number of independent parameters (D) and the number of best selected models of SERE-homogeneous data for different sample sizes.

		N = 1000			N = 2000			N = 5000		
Model	D	PAIC	SDV	Best model	PAIC	SDV	Best model	PAIC	SDV	Best model
a	51	574	28	0	678	53	0	1014	56	0
b	52	564	29	0	657	51	0	957	59	0
c	54	491	13	11	493	16	3	536	16	0
d	55	487	11	14	484	12	22	505	8	25

As already mentioned in Section 4.3, even with small sample sizes the decision whether an item shows DIF or not can be made almost as well as with large sample sizes. With a sample size of 1000 subjects for 14 out of 25 replications the true model (d) was selected as the best. However, the lowest chance of making a wrong decision is found in the sample size of 5000 subjects: all 25 replications selected model d as the best.

4.5.3 The influence of the test length on DIF detection and parameter estimates

In this section we will examine whether DIF can still be found if the test length decreases. For this examination three data sets were generated: one consisted of eight items (i.e. the nine items of Table 4.1, except Item 4), the second consisted of six items (Items 2,3,5,6,7,8) and the third consisted of the Items 2, 5, 6 and 8. Please remember that Item 8 shows DIF in the latent response, whereas Item 2 shows DIF in the attraction parameters. Furthermore, the sample sizes were equal to 1000 and 25 replications were made. Finally, the parameters of the models mentioned in Section 4.5.1 (i.e. Model a through d) were estimated. To compare the three data sets, the consistent pseudo Akaike's information criterion values (PAIC) and the number of independent parameters (D) for each of the four models are presented in Table 4.4. And just as in Table 4.3 the number of occasions in which the particular model had the lowest PAIC value are also given.

Table 4.4
Mean of the consistent pseudo Akaike's information criteria (PAIC), the number of independent parameters (D) and the number of lowest PAIC values of SERE-homogeneous data for different numbers of items.

	4 items			6 items			8 items		
Model	D	PAIC	Best model	D	PAIC	Best model	D	PAIC	Best model
a	51	574	0	68	1492	0	115	1586	0
b	52	564	0	69	1397	0	116	1560	0
c	54	491	11	71	667	3	118	1370	2
d	55	487	14	72	598	22	119	1350	23

From the results of Table 4.4 we may conclude that DIF detection is better when there are more unbiased items in the test. The comment by Shepard, Camilli and Williams (1985) that too many biased items in the test would harm the stability of detecting DIF might therefore be valid for the SERE model. In the first set of data fifty percent of the items were biased, but with a sample size of 1000 subjects all biased items were detected in only 14 of the 25 replications, whereas for the second and third set of data all biased items were detected in almost all 25 replications.

In order to answer the question whether the deviations between the parameter values of the estimated parameters and the true parameters decrease when the number of items in the test increases, in Table 4.5 the mean of the estimated attraction parameters of the right alternatives and the estimated difficulty parameters for the three sets of data of the first group are depicted.

Table 4.5
Mean and standard deviations (SDV) of the attractiveness of the correct alternative and the item difficulties for the first group of SERE-homogeneous data in model d for different numbers of items. (N = 1000)

		4 items		6 items		8 items	
Item	True	Estimated	SDV	Estimated	SDV	Estimated	SDV
				Attraction parameters			
2	.25	.1195	.0221	.1215	.0207	.1224	.0196
5	.25	.1474	.0070	.1466	.0070	.1470	.0071
6	.25	.2022	.0098	.2002	.0098	.2014	.0100
8	.25	.4172	.0186	.4179	.0190	.4228	.0189
				Item difficulties			
2	1.5	1.084	.0703	1.091	.0721	1.080	.0688
5	0.0	0.000	-	0.000	-	0.000	-
6	-0.5	-0.290	.0608	-0.291	.0599	-0.288	.0569
8	-1.5	-0.701	.0922	-0.712	.0885	-0.705	.0832

This table shows that in the three sets of data the differences between the estimated values of the parameters were not very large. When we compare the three sets of data, the standard deviations of the estimates of the item difficulties, except for Item 2, seem to have decreased with increasing test length. This trend is not very clear for the attraction parameters. For some items the standard deviations of the estimates of the attractiveness of the right alternative decreased and for other items the standard deviations increased.

For the second group the conclusions are not different from those about the first group. Therefore they are not given.

4.5.4 The simultaneous detection of DIF in the latent response and DIF in the attraction parameters

In the fourth part of the simulation study the attention was focused on the combination of the two types of DIF: Is it possible to detect an item which shows DIF in the latent response, but shows also DIF in the attraction parameters? In order to answer this question, a data set was generated for Items 1, 3, 4, 5, 6, 7, and 9 of Table 4.1. Please note that Item 4 is the only item which shows DIF both in the latent response and in the attraction parameters. Again the data set was generated for two groups of 1000 subjects, and 25 replications were made. This time, the pseudo-likelihood statistics were calculated for the models in which Item 4 shows (e) no DIF, (f) DIF in the latent response, (g) DIF in the attraction parameters, or (h) DIF in the latent response and DIF in the attraction parameters. The results are presented in the Table 4.6, whereby the contents of the columns are similar to those of Tables 4.3 and 4.4.

Table 4.6
Mean of the consistent pseudo Akaike's information criteria (PAIC) with their standard deviation (SDV), the number of independent parameters (D) and the number of lowest PAIC values of SERE-homogeneous data.

Model	D	PAIC	SDV	Best model
e	153	1685	59	0
f	154	1549	45	0
g	156	1456	29	0
h	157	1447	30	25

Table 4.6 indicates that it is really possible to detect items which show DIF in the latent response as well as DIF in the attraction parameters. For all replications the model in which item 4 was the only item that showed both types of DIF (i.e. the true model) was selected as the best model in comparison with models in which item 4 shows no DIF or only one type of DIF. However, when the sample size was 5000, in only 3 replications the true model was selected as the best. In the other 23 replications the model in which item 4 only shows DIF in the attraction parameters (i.e. model g) was selected as the best. In view of the mean of the PAIC values for the two models g and h (1488 and 1491, respectively), the reason for the discrepancy between the two sample sizes might be the choice of the pseudo Akaike's information criterion. Generally, the consistent pseudo Akaike's information criterion tends to lead to simpler models than the pseudo Akaike's information criterion does, which happened for the sample size of 5000. When the pseudo Akaike's information criterion model selection method was used, then for both sample sizes model h (i.e. the true model) had always been selected as the best model.

4.5.5 Small-sample behavior of simultaneous detection of DIF in the latent response and DIF in the attraction parameters

In the preceding sections, the research questions were all restricted to (1) situations in which the sample size was relatively small or large, (2) situations in which only one item in the item set shows DIF in the latent response and only one other item shows DIF in the attraction parameters, and (3) situations in which only one item in the item set shows both types of DIF. In real-life situations, however, extremely small sample size of 100 or 250 subjects are commonly used. Moreover, in real-life situations more than one item may show one of the two types of DIF or may show both types of DIF.

We have seen that in a situation where in a set of data one item shows one type of DIF and another item shows the other type of DIF, it is possible to detect both biased items. We have also seen that it is even possible to detect items that show both types of DIF. Would we have find the same results if there were more biased items in the test? And can we still found the items which shows DIF if the sample size is extremely small?

Table 4.7

Mean of the consistent pseudo Akaike's information criteria (PAIC) with their standard deviation (SDV), the number of independent parameters (D) and the number of lowest PAIC values of SERE-homogeneous data for different sample sizes.

		N = 100			N = 250			N = 500		
Model	D	PAIC	SDV	Best model	PAIC	SDV	Best model	PAIC	SDV	Best model
0000	143	2668	113	0	2611	95	0	2285	89	0
0001	144	2660	111	0	2588	91	0	2264	90	0
0010	144	2655	108	0	2571	90	0	2211	95	0
0011	145	2651	107	0	2556	85	0	2198	95	0
0100	146	2619	102	0	2536	71	0	2155	71	0
0101	147	2612	100	0	2515	64	0	2140	73	0
0110	147	2615	98	1	2519	70	0	2135	73	0
0111	148	2611	96	0	2503	63	0	2122	74	0
1000	146	2646	111	0	2541	79	0	2172	72	0
1001	147	2639	111	1	2519	72	0	2153	74	0
1010	147	2633	107	1	2499	75	0	2090	76	0
1011	148	2628	106	1	2484	66	0	2080	77	1
1100	149	2597	98	0	2464	63	0	2037	47	0
1101	150	2591	97	5	2445	51	2	2025	49	0
1110	150	2593	94	2	2446	63	6	2014	47	2
1111	151	2589	92	10	2432	50	17	2004	48	22

Note. For the sample size of 100 four replications of the generated set of data were omitted from the simulation study because of problems during the estimation process.

In order to answer these two questions two sets of data were generated, one for each subgroup. Each set of data consists of all nine items as defined in Table 4.1. Furthermore, sample sizes of 100, 250 and 500 were chosen, and for every sample size 25 replications were made. Please note that the Items 2 and 4 showed DIF in the attraction parameters and Items 4 and 8 showed DIF in the latent response. The parameiers of 16 models were estimated. In each model it was postulated whether or not Items 2 and 4 show DIF in the attraction parameters and whether or not Items 4 and 8 show DIF in the latent response. In the Tables 4.7 and 4.8 the results are presented.

In these tables the models will be denoted by a chain of four digits: zero or one. The first digit in the chain defines whether it was postulated that Item 2 shows DIF in the attraction parameters, where a zero means "No" and a one means "Yes". In the same way, the second, third and fourth digit declares whether Item 4 shows DIF in the attraction parameters, Item 4 shows DIF in the latent response or Item 8 shows DIF in the latent response. The chain 1010, for example, defines a model in which item 2 shows DIF in the attraction parameters and item 4 shows DIF in the latent response, whereas the chain 1101 defines a model in which items 2 and 4 show DIF in the attraction parameters and item 8 shows DIF in the latent response.

In Table 4.7 the values of the consistent pseudo Akaike's information criterion (PAIC) of the 16 models are given for the three different sample sizes. Furthermore, the standard deviation of these PAIC values and the number of occasions in which the particular model had the lowest PAIC value, are given.

From the results of Table 4.7 we may conclude that even for situations in which the sample size is extremely small and there is more than one item that showed DIF, it is possible to detect all these items. With the smallest sample size (i.e. 100) in 10 of the 21 replications the true model was selected as the best, whereas with a sample of 250 or 500 subjects the number of

Table 4.8

Mean and standard deviations of the attractiveness of the correct alternative and item difficulties of SERE-homogeneous data in the true postulated model (1111) for different sample sizes

		N = 100		N = 250		N = 500	
Item	True	Estimated	SDV	Estimated	SDV	Estimated	SDV
				Attraction parameters			
2	.25	.1137	.0351	.0928	.0146	.1078	.0176
5	.25	.2742	.0401	.1934	.0231	.1637	.0110
6	.25	.3671	.0614	.2749	.0173	.2265	.0184
8	.25	.6940	.0680	.5682	.0475	.4832	.0276
				Item difficulties			
2	1.5	0.711	.1120	0.876	.1108	0.978	.0858
5	0.0	0.000	-	0.000	-	0.000	-
6	-0.5	-0.175	.1170	-0.255	.0999	-0.262	.0836
8	-1.5	-0.411	.1042	-0.498	.1461	-0.618	.0882

replications in which the true model was selected as the best was equal to 17 and 22, respectively.

In order to examine the deviations between the values of the parameters of the estimated model 1111 and those of the true model, in Table 4.8, for each sample size, the true and the mean of the estimated attractiveness of the right alternatives and the estimated difficulties are given for the Items 2, 5, 6 and 8 and for the first group. Furthermore, in the column labelled "SDV" the values of the standard deviates of the estimates are given.

Just as in the case of the relatively large sample sizes (i.e. 1000, 2000 and 5000), this part of the simulation study shows that the estimated attractiveness of the correct alternative was biased, but that the difference between the true and the estimated difficulties decreased with increasing sample size. The inconsistency of the estimators of the attractiveness of the correct alternative was again due to the instability of the parameter estimates of three-parameter logistics models.

Here, too, the conclusions about the second group were no different from those about the first group. Therefore they are also not given.

4.6 CONCLUSIONS AND DISCUSSION

This chapter dealt with the question whether DIF can be found with the SERE model and how the values of the parameters of the estimated SERE model differ from those of the original model.

In the first place, from the results of this simulation study we may conclude that despite the trade-off between the difficulty parameters and the attraction parameters, the difference between the true and the estimated difficulty decreased with increasing sample sizes or increasing test lengths. For the attractiveness of the correct alternative this relation between the sample size or test length with the deviation between the values of the parameters of the estimated model and the original model could not be found.

Secondly, the simulation study showed that with (extremely) small sample sizes DIF could still be detected, but that the chance of the detection of DIF increased when the sample size increased. Evidence that the test length has an effect on DIF detection could be found: there were indications that too many biased items harmed the stability of detecting DIF.

Finally, one of the main reasons why the SERE model was developed was for the examination of items not only for DIF due to item difficulty but also due to alternative attractiveness. From the last two parts of the simulation study we may conclude that with the

SERE model a distinction can be made between both types of DIF and that items can be detected which show both types of DIF.

Chapter 5

GENERALIZATIONS OF THE SOLUTION-ERROR RESPONSE-ERROR MODEL*

5.1 ABSTRACT

In the last decade, several efforts have been made to relate item response theory (IRT) models to latent class analysis (LCA) models. One of these efforts is the solution-error response-error (SERE) model; a LCA model in which the structure of the latent class probabilities is explained with a one-dimensional loglinear Rasch model.

In this chapter the SERE model will be generalized to models for polytomously scored latent states that may be explained by a multidimensional latent space.

5.2 INTRODUCTION

For the measurement of individual differences, a distinction can be made between measurements on a discrete qualitative latent trait and measurements on a continuous quantitative scale. The latent class analysis (LCA) model, in which the assumption is that subjects belong to different latent classes, is an example of the former (Bartholomew, 1987; Lazarsfeld & Henry, 1968; Mooijaart, 1978). Whereas the item response models (IRT) is an example of the latter. Some well-known examples of IRT models are the Rasch (1960/1980) model and the two- and three-parameter-logistic or normal ogive models (Lord, 1980; Lord & Novick, 1968). In the last decade several efforts have been made to relate IRT models to LCA models (Bock & Aitkin, 1981; Dayton & Macready, 1980; Formann, 1985; Kelderman, 1988, 1989; Kelderman & Macready, 1990; Mislevy & Verhelst, 1990; Yamamoto, 1987, 1988). In this chapter one of these efforts will be discussed: the solution-error response-error (SERE) model of Kelderman (1988).

---

* The authors P. Westers and H. Kelderman would like to thank Mary E.Lunz and Barbara M. Castleberry of the American Society of Clinical Pathologist (ASCP) for providing empirical data used in the example in this chapter.

In the SERE model a distinction is made between a "Know" state in which the subject has a complete knowledge of the answer, and a "Don't know" state. The probability that the subject is in the "Know" state is assumed to be governed by the Rasch model. Furthermore, the assumption is that whether or not the subject is in the "Know" state, (s)he will choose the most attractive alternative, in which the attractiveness may be dissimilar for different alternatives, including the correct one. The SERE model can be formulated as an (incomplete) LCA model, in which each latent class corresponds with an idealized response pattern. The relations between these idealized responses are explained by the loglinear version of the Rasch model (Cressie & Holland, 1983; Duncan, 1984; Kelderman, 1984; Tjur, 1982).

All SERE models considered in Kelderman (1988) deal with a one-dimensional continuous latent trait. In many testing situations, however, we may have to deal with a two - or more - dimensional latent space. Consider, for example, a version of the American Society of Clinical Pathologist (ASCP) Microbiology Test. In Appendix A.4 some items of this test are presented. Content experts have hypothesized that although each item of this ASCP test has one correct alternative, incorrect responses might often be chosen after cognitive activities similar to those necessary to arrive at the correct response. They further presumed that "Applying Knowledge", "Selecting Action", "Calculating", "Correlating Data" and "Evaluating Problem" are the cognitive processes involved in answering the items. For instance, they assumed that for item 11 of Appendix A.4 the correct answer (d) involved two applications of knowledge, whereas answer c involved only one. In order to give the correct answer c on item 20 they assumed that the subject had to use the cognitive process "Evaluating Problem" twice and the cognitive processes "Applying Knowledge" and "Selecting Action" once.

So, in general, the production of one answer may require quite another ability from the examinee than the production of another. Or some responses may require the repeated application of an ability, whereas others may require only a single application of the same ability. In this chapter the SERE model will be generalized to models for polytomously scored latent states that may be explained by a multidimensional latent space. Maximum likelihood estimates of the parameters of this generalized SERE (GSERE) model can be obtained by solving the likelihood equations by the iterative proportional fitting (IPF) algorithm of Goodman (1974b).

The GSERE model will be formulated below. The estimation method and goodness-of-fit tests are described, and the question of identifiability is discussed.

5.3 THE GENERALIZED SOLUTION-ERROR RESPONSE-ERROR MODEL

Let us suppose that each subject, randomly drawn from a population of subjects, responds to k test items, in which the answer to item j may be any of the r_j responses, denoted by y_j ($y_j=1,\dots,r_j$). Let x_j ($x_j=0,\dots,s_j$) indicate the latent state of the subject. For example, all the items of the ASCP Microbiology Test have four possible responses (i.e., $r_j=4$) and may have three latent states: "Don't know", "Partial knowledge" and "Complete knowledge". We will assume that when the subject is in the "Don't know" state ($x_j=0$), (s)he will choose one of the alternatives. If the subject is in the "Partial knowledge" state ($x_j=1$), (s)he will choose one of the alternatives that might be correct in view of the subject's partial knowledge of the answer. If the subject is in the "Complete knowledge" state ($x_j=2$), (s)he will choose the correct alternative. The random variables with values y_j and x_j are denoted by Y_j and X_j ($j=1,\dots,k$). The relationship between the latent state x_j and the observed response y_j is described by the conditional probability

$$\text{(5.1)} \qquad \Phi^{X_jY_j}_{x_jy_j} \equiv P(Y_j=y_j|X_j=x_j)\ ,$$

This conditional probability will be referred to as the attraction parameter of item j.

In the generalized solution-error response-error (GSERE) model we assume that the latent states are not governed by the Rasch (1960/1980) model, but by the more general multidimensional polytomous latent trait model (MPLT) by Kelderman and Rijkes (in press). In the MPLT model the assumption is that the subject must perform certain cognitive operations to produce a latent score x_j on item j. See for instance the example in the previous section. Each operation depends on a certain proficiency on a latent trait. Let $B_{jq}(x)$ be a non-negative weight associated with the dependence of response x on item j on the latent trait q. Furthermore, let $\delta_{jq}(x)$ be the difficulty parameter of the response x on item j related to latent trait q ($q=1,\dots,v$), θ_q be a value of the subject on the latent ability continuum and $\theta = (\theta_1,\dots,\theta_v)$ be the vector of ability values. The probability that the subject has a response x_j on item j can then be written as (Kelderman & Rijkes, in press)

$$\text{(5.2)} \qquad P(x_j|\theta) = \frac{\exp\{\sum_q (\theta_q - \delta_{jq}(x_j))B_{jq}(x_j)\}}{\sum_z \exp\{\sum_q (\theta_q - \delta_{jq}(z))B_{jq}(z)\}}$$

Assuming local independence of X_j and Y_j given the latent trait vector θ and X_j, respectively, the probability of choosing response y_j is equal to

$$P(y_j|\theta) = \sum_{x_j} P(y_j|x_j, \theta)\, P(x_j|\theta) \tag{5.3}$$

$$= \sum_{x_j} \{ \Phi^{X_jY_j}_{x_j y_j} \} \exp\{ \sum_q (\theta_q - \delta_{jq}(x_j))B_{jq}(x_j) \}\, C(\theta,\delta_j)^{-1}$$

in which $\delta_j = (\delta_{j1},...,\delta_{jv})$ and

$$C(\theta,\delta_j) = \sum_z \exp\{ \sum_q (\theta_q - \delta_{jq}(z))B_{jq}(z) \} .$$

As Kelderman and Rijkes (in press) have shown with the specification of the scoring weights $B_{jq}(.)$, different models can be chosen for the dependence of the latent states on the latent traits. To illustrate the main idea of this chapter, one specific MPLT model will be considered below: the multidimensional partial credit (MPCM) model. It may, however, be clear that the contents of this chapter is also valid for other kinds of MPLT models.

The scoring weights for a MPCM model, in which each step depends on a different latent trait, are depicted in Figure 5.1(a).

		q	
		1	2
x	0	0	0
	1	1	0
	2	1	1

(a)

		q
		1
x	0	0
	1	1
	2	2

(b)

Figure 5.1
Scoring weights for the one- and two-dimensional partial credit model.

The "Complete knowledge" state (x=2) has scoring weight $B_{j1}(2) = 1$ on the first trait and scoring weight $B_{j2}(2) = 1$ on the second. The "Partial knowledge" state (x=1) has scoring weight $B_{j1}(1) = 1$ on the first trait, whereas the "Don't know" state (x=0) has scoring weights zero. With the use of (5.2) and the scoring weights $B_{jq}(.)$ of Figure 5.1(a), the probability that the subject has a latent state x_j on item j can be written as

$$P(x_j|\theta) = \frac{\exp\{\sum_{q=1}^{x_j} (\theta_q - \delta_{jq}(x_j))\}}{\sum_z \exp\{\sum_{q=1}^{z} (\theta_q - \delta_{jq}(z))\}} . \qquad (5.4)$$

Adding a constant c to $\delta_{jq}(l)$ and subtracting it from $\delta_{jq}(l')$ $(1 \leq l \neq l' \leq x_j)$ does not change the model in (5.4). By setting the difficulty parameters of the same response equal to each other (i.e. $\delta_{jq}(x_j) = \delta_{jq}$ for $x_j = 0,1,2$ and all q), this indeterminacy can be removed (Kelderman & Rijkes, in press). From (5.4) and the assumption of local independence of the X_j's given the latent trait vector θ, it follows that the simultaneous distribution of X given θ is

$$P(\mathbf{x}|\theta) = \exp\{\sum_q (\theta_q t_q - \sum_j B_{jq}(x_j)\delta_{jq})\} \prod_j C(\theta,\delta_j)^{-1} \qquad (5.5)$$

in which

$$C(\theta,\delta_j) = \sum_z \exp\{ \sum_{q=1}^{z} (\theta_q - \delta_{jq})\}$$

and

$$t_q = \sum_j B_{jq}(x_j) \qquad q=1,\dots,v.$$

Just like all other MPLT models the MPCM model is an exponential family model and $\mathbf{t}=(t_1,\dots,t_v)$ is a sufficient statistic for the latent trait vector θ (Kelderman & Rijkes, in press).

Let $\Sigma_{\mathbf{x}}$ mean the summation over all possible latent state patterns $\mathbf{x}=(x_1,\dots,x_k)$. With the use of (5.1), (5.5) and the assumption of local independence of the observed responses y_j, the marginal probability of $\mathbf{y}$ given θ can be written as

$$P(\mathbf{y}|\theta) = \sum_{\mathbf{x}} P(\mathbf{y}|\mathbf{x}, \theta)\, P(\mathbf{x}|\theta) \qquad (5.6)$$

$$= \sum_{\mathbf{x}} \{ \prod_j \Phi_{x_j y_j}^{X_j Y_j} \} \exp\{ \sum_q (\theta_q t_q - \sum_j B_{jq}(x_j)\delta_{jq})\} \prod_j C(\theta,\delta_j)^{-1}.$$

If the terms of (5.4) are arranged and if $F(\boldsymbol{\theta})$ is the distribution function of the latent ability vector $\boldsymbol{\theta}$, the marginal probability of the observed responses **y** can be written as an incomplete LCA model in the sense of Haberman (1979)

$$P(\mathbf{y}) = \sum_{\mathbf{x}} \Phi^{T}_{\mathbf{t}} \Phi^{X_1}_{x_1} \ldots \Phi^{X_k}_{x_k} \Phi^{X_1 Y_1}_{x_1 y_1} \ldots \Phi^{X_k Y_k}_{x_k y_k} \tag{5.7}$$

with

$$\Phi^{T}_{\mathbf{t}} = \int \exp(\sum_{q} \theta_q t_q) \prod_{j} C(\boldsymbol{\theta}, \boldsymbol{\delta}_j)^{-1} dF(\boldsymbol{\theta})$$

and

$$\Phi^{X_j}_{x} = \exp(-\sum_{q=1}^{x} \delta_{jq}), \qquad (j=1,\ldots,k).$$

In this model each value of the latent state vector **x** represents a latent class. Maximum likelihood estimates of the parameters of the GSERE model can be obtained by solving the likelihood equations with the iterative proportional fitting (IPF) algorithm (Bartholomew, 1987; Goodman, 1974b; Haberman, 1979; Hagenaars, 1990). The overall goodness-of-fit of a model can be tested by the Pearson statistic or the likelihood-ratio statistic (see Haberman, 1979). Together with the question of identifiability, these two issues will be discussed in the next sections. But some applications of the GSERE model will be discussed first.

5.4 APPLICATIONS OF THE GSERE MODEL

In the previous section a GSERE model was formulated in which the parameters of the model were unrestricted, except for the usual restrictions pertaining to probabilities and conditional probabilities. In this section, we will discuss how the GSERE model can be modified in order to make it suitable for special applications.

Generally, for each specific GSERE model we may define the weights $B_{jq}(.)$ and certain constraints on the attraction parameters for each item j. The choices of the weights may depend on the required latent trait abilities for the correct response. For example, the item "20-5-6=?"

requires two subtraction operations for the correct response, so that we can choose the one-dimensional partial credit model as depicted in Figure 5.1(b). But for the item "√(169-25)=?", in which the two abilities are subtraction and taking the square root, we can choose the two-dimensional partial credit model as depicted in Figure 5.1(a). In Kelderman and Rijkes (in press) other possible choices of the scoring weights for the dependence of the latent states on latent traits are discussed.

		y	
		1	2
x	0	0	1
	1	1	0

(a)

			y		
		1	2	3	4
x	0	α_1	α_2	α_3	α_4
	1	1	0	0	0

(b)

			y		
		1	2	3	4
x	0	α_1	α_2	α_3	α_4
	1	β_1	0	β_2	0

(c)

		y	
		1	2
x	0	0	1
	1	1-β	β

(d)

			y		
		1	2	3	4
	0	α_1	α_2	α_3	α_4
x	1	β_1	0	β_2	0
	2	1	0	0	0

(e)

Figure 5.2
Examples of specifications of the attraction parameters

By specifying the attraction parameters (5.1) as free, equal to each other or fixed to a certain value, a particular GSERE model is specified (Kelderman, 1988; Westers & Kelderman, 1992). In Figure 5.2 some examples of the attraction parameters for the GSERE model are depicted. Figure 5.2(a) describes the situation of a perfect response process; the subject answers the item correctly (y=1) if (s)he can solve the problem (x=1) and gives a wrong answer (y_j=2) if (s)he cannot solve the problem (x=0). A case in which the items are not necessarily answered incorrectly if the subject cannot solve the problem is when there are multiple-choice items. If the subject doesn't know the answer, (s)he will guess the most attractive alternative. The attraction parameters for this situation are depicted in Figure 5.2(b). In Figure 5.2(c) the situation is depicted for the case in which more than one alternative is correct. The general assumption is that if the subject can solve the problem formulated by the item, (s)he will give the correct answer. But a subject may fail to produce a correct answer, even if (s)he was able to solve the problem. On the other hand, if (s)he is not able to solve the problem it is impossible to produce the correct answer. Such a situation is depicted in Figure 5.2(d), in which β is the so-called omission error. In the case of the MPCM model we may assume that the attraction parameters are specified as depicted in Figure 5.2(e).

5.5 IDENTIFIABILITY

Whether the maximum likelihood estimates of the parameters of the GSERE model are unique depends on the identifiability of the model. A necessary condition for identifiability is, of course, that the number of independent parameters to be estimated does not exceed the number of cell frequencies minus one (i.e. $(\Pi\, r_j)-1$). Furthermore, if the MPLT model is not (locally) identifiable, the GSERE model is not (locally) identifiable either.

Generally, the parameters in a MPLT model ought to be restricted in order to obtain an identifiable model. Therefore, in the paper of Kelderman and Rijkes (in press) conditions are formulated which ensures that the difficulty parameters of the MPLT model are not linearly dependent upon each other and upon the proportionality constants.

Since the (G)SERE model can be formulated as an (incomplete) LCA model, Goodman's (1974a) sufficient condition for identifiability can be used for the identifiability of the GSERE model. Let M be the matrix consisting of the derivatives of the function (5.7) with respect to the parameters of the GSERE model. The number of rows of the matrix M is equal to $(\Pi\, r_j)-1$ and the number of columns is equal to the number of parameters of the GSERE model. By direct extensions of a standard result about Jacobians, the GSERE model will be locally identifiable if the rank of the matrix M is equal to the number of columns. The rank of the matrix M can be evaluated by numerical methods.

When the parameters of the GSERE model are not identifiable, various kinds of restrictions can be imposed upon the attraction parameters and/or the item parameters in order to make the GSERE model identifiable. The attraction parameters, for instance, may be equated with each other or with a constant.

Unidentifiability can be discovered by estimating the parameters a second time, this time with the use of different initial estimates. In the case of unidentifiability both runs will give different parameter estimates.

5.6 ESTIMATION METHOD

Let $m_{\mathbf{x}yt}$ be the expected number of subjects with latent state **x**, observed response y and sum-score t. As Kelderman (1988), and Westers and Kelderman (1992) have shown, the parameters of the SERE model can be estimated by applying the iterative proportional fitting (IPF) algorithm. For the GSERE the IPF algorithm can also be used. One of the differences in the SERE model is that, depending on the number of the latent state categories, the number of latent classes may be quite large. Since the number of attraction parameters depends on the number of latent states categories and the number of item response categories, this number may also be quite large.

The maximum likelihood estimates of the parameters of the GSERE model can be obtained by solving the likelihood equations by a two-step algorithm. In the first step of each iteration (i.e. the outer iteration), the attractiveness of the alternatives and the expected frequency distribution of the latent classes will be estimated. For the GSERE model the first step is similar to the first step of the estimation method for the parameters in the LCA model (Goodman, 1974b). In the second step of each iteration (i.e. the inner iteration) the estimated expected distribution of the latent classes is fitted to the postulated loglinear model. From this distribution the difficulty parameters can be estimated.

5.6.1 Outer Iteration

As indicated before, the GSERE model can be formulated as a LCA model, in which each latent class represents a latent state vector **x**. Let

$$P(\mathbf{y}) = \sum_{\mathbf{x}} P(\mathbf{x})\, P(y_1|\mathbf{x}) \ldots P(y_k|\mathbf{x}) \tag{5.8}$$

in which P(**x**) is the probability of getting latent state vector **x** and $P(y_j|\mathbf{x})$ is the conditional probability of choosing response y_j given the latent state vector **x**. The model in (5.8) is a LCA model in the sense of Goodman (1974b), which means that the IPF algorithm for the general latent structure model, in which the parameters of the model are unrestricted, can be used for the estimation of the expected frequency distribution of the latent classes and the attractiveness of the alternatives.

As mentioned before, in the GSERE model we assumed that the observed response y_j depends only on the latent state x_j; therefore the conditional probabilities $P(y_j|\mathbf{x})$ are restricted in the following manner

$$P(y_j|\mathbf{x}) = P(y_j|\mathbf{x}') = \Phi^{X_jY_j}_{x_j\,y_j}$$

for all latent state vectors $\mathbf{x}$ and $\mathbf{x}'$ with components $x'_j = x_j$. The Φ-parameters can be obtained from a weighted average of the estimates $P(y_j|\mathbf{x})$ obtained from the IPF algorithm, with weights proportional to $P(\mathbf{x})$ (Goodman, 1974b). So

$$\Phi^{X_jY_j}_{x_j\,y_j} = \{ \sum_{\mathbf{x}'} P(\mathbf{x})\, P(y_j|\mathbf{x}) \} \,/\, \{ \sum_{\mathbf{x}'} P(\mathbf{x}) \} ,$$

in which $\Sigma_{\mathbf{x}'}$ is the summation over all latent state vectors $\mathbf{X}=\mathbf{x}$ with $X_j=x_j$.

5.6.2 Inner Iteration

As assumed in Section 5.3, the latent probabilities $P(\mathbf{x})$ are restricted in such a way that they comply with a MPCM model. Knowing that the MPCM model is an exponential family model and that the sum-score $\mathbf{t}$ is a sufficient statistic for the latent ability parameter θ (Kelderman & Rijkes, in press), the conditional distribution of $\mathbf{X}$ given $\mathbf{t}$ is

$$P(\mathbf{x}|\mathbf{t}) = \exp \{ \sum_j \varphi_j(x_j) \} \,/\, g(\mathbf{t},\varphi),$$

with

$$g(\mathbf{t},\varphi) = \sum_{\mathbf{x}|\mathbf{t}} \exp\{ \Sigma\, \varphi_j(x_j)\},$$

in which $\Sigma_{\mathbf{x}|\mathbf{t}}$ is the summation over those values of the latent state vector $\mathbf{x}$ for which $(\Sigma\, B_{j1}(\mathbf{x}),\dots,\Sigma\, B_{jv}(\mathbf{x}))$ is equal to $\mathbf{t}$, and the vector $\varphi = (\varphi_1(x_1),\dots,\varphi_k(x_k))$ denotes the weighted sums over latent traits of the difficulty parameters

$$\varphi_j(x_j) = - \sum_{q=1}^{v} \delta_{jq}(x_j)\, B_{jq}(x_j).$$

If $m_{\mathbf{xt}}$ is the estimated expected number of subjects with latent state **x** and sum-score **t**, we have

$$\log m_{\mathbf{xt}} = \sigma_{\mathbf{t}} + \sum_{j} \varphi_j(x_j) \tag{5.9}$$

in which $\sigma_{\mathbf{t}} = -\log(\, g(\mathbf{t},\varphi)/n_{\mathbf{t}})$ is a fixed normalizing constant, $n_{\mathbf{t}}$ is the number of subjects with sum-score **t**. For the MPCM model the parameters $\varphi_j(x_j)$ are specified by $\varphi_j(1) = \delta_{j1}$ and $\varphi_j(2) = -\delta_{j1} - \delta_{j2}$, respectively.

The model in (5.9) is a quasi-loglinear model for an incomplete item response 1 x...x item response k x score 1 x...x score v contingency table. The table is incomplete since for certain given values of **X** only one **t** is possible. Maximum likelihood estimates can be obtained by solving the likelihood equations of the MPCM model. These equations can be solved by IPF (Kelderman & Rijkes, in press). The latent probabilities P(**x**) are then adjusted to these maximum likelihood estimates. In this way new latent probabilities P(**x**) are obtained that comply with the postulated MPLT model and are used again in the next outer iteration.

5.7 GOODNESS-OF-FIT TEST

The overall goodness of fit of the GSERE model can be tested by the Pearson statistic or the likelihood-ratio test statistic. Both statistics are asymptotically distributed as chi square with degrees of freedom equal to the difference between the number of cells in the observed contingency table minus one and the number of estimable parameters. If the expected counts, however, become too small, the approximation of the distribution of the goodness-of-fit statistics by a chi-square distribution will be bad (Haberman, 1988; Koehler, 1977, 1986; Lancaster, 1961; Larnz, 1978).

With the use of the difference in the likelihood-ratio test statistics for two models (Bishop, Fienberg, & Holland, 1975; Rao, 1973), it can be tested whether an alternative model gives a significant improvement in fit over a special case of this alternative model.

If the two GSERE models are not proper subsets of each other, then Akaike's (1977, 1987) information criterion (AIC) or Raftery's (1986a, 1986b) Bayesian information criterion (BIC) can be used. AIC is defined as

$$\mathrm{AIC} = G^2 - 2\,d$$

in which G^2 is the likelihood-ratio test statistic and d is the number of independent parameters in the GSERE model. The BIC index has ln n (i.e. n is the sample size) instead of 2, but is otherwise identical. For both indices, the first term is a measure of badness of fit, whereas the second term is a correction for overfitting due to the increasing bias in G^2 as the number of parameters in the SERE model increases. The GSERE model with the minimum AIC or BIC value will be chosen as the best fitting model. Computer programs by Hagenaars and Luijkx (LCAG, 1990) and Kelderman and Steen (LOGIMO, 1988) can be used to fit the model.

5.8 AN EMPIRICAL EXAMPLE

For numerical reasons the GSERE model is still difficult to apply routinely in large testing programs. Not the number of parameters of the model causes the problems, but the number of latent classes and the tables of observed and expected counts become too large for computer storage. One solution to this problem may be the use of the division-by-item (DBI) principle from Chapter 3. In this chapter a maximum likelihood estimation method for the one-dimensional SERE model for a large set of items was described. This method was based on the division of the whole item set in several subsets. The computational problem boils down to the simultaneous estimation of the parameters of a set of smaller SERE models. This could be done because one of the properties of the SERE model was that the model could be collapsed. Since this property is also valid for the GSERE model (Appendix A.5), we can use the DBI-principle for the estimation of the parameters of the GSERE model. Before the model can be widely applied, however, further research is required to reduce the amount of computer storage. But if the generalization for the multidimensional latent space is ignored, the parameters of the GSERE model can still be estimated for a small number of items, with the combined use of the programs LCAG and LOGIMO. In this section this will be demonstrated with an example.

Table 5.1
Hypothesized weights of the ASCP Medical Laboratory Test Items for the cognitive process "Applying Knowledge".

Item	Scoring weights				Correct answer
	A	B	C	D	
1			2	1	C
2	1		2		C
3	2	1	1		A
4	2	1	1		A

The authors were allowed to analyse four one-dimensional four-choice items from a protected data base of the ASCP Medical Laboratory Test. ASCP produces tests for certification of paramedical personnel. With the items in the Medical Laboratory Test the ability to perform medical laboratory tests will be measured. The test score is obtained by adding the number of correct items.

Content experts from ASCP have hypothesized that the cognitive process "Applying Knowledge" was involved in answering these four items. According to the assumptions of the content experts, we postulated that there are three latent states: "Don't know", "Partial knowledge" and "Complete knowledge", with scoring weights equal to zero, one and two, respectively. This means that the latent states are assumed to be governed by the one-dimensional partial credit model (OPCM) as depicted in Figure 5.1(b). Table 5.1 shows the hypothesized weights that content experts gave for each of the item responses on the cognitive process "Applying Knowledge". These hypothesized scoring weights are translated into specifications of the attraction parameters, which are depicted in Figure 5.3, in which x=0

			y	
	A	B	C	D
0	α_1	α_2	α_3	α_4
x 1	0	0	β_1	β_2
2	0	0	1	0

(Item 1)

			y	
	A	B	C	D
0	α_1	α_2	α_3	α_4
x 1	β_1	0	β_2	0
2	0	0	1	0

(Item 2)

			y	
	A	B	C	D
0	α_1	α_2	α_3	α_4
x 1	β_1	β_2	β_3	0
2	1	0	0	0

(Item 3)

			y	
	A	B	C	D
0	α_1	α_2	α_3	α_4
x 1	β_1	β_2	β_3	0
2	1	0	0	0

(Item 4)

Figure 5.3
Specifications of the attraction parameters of the four items from the ASCP Medical Laboratory Test for the hypothesized model H.

indicates the "Don't know" state, x=1 the "Partial knowledge" state and x=3 the "Complete knowledge" state.

			y		
		A	B	C	D
	0	α_1	α_2	α_3	α_4
x	1	β_1	β_2	β_3	β_4
	2	γ_1	γ_2	γ_3	γ_4

(A_1)

			y		
		A	B	C	D
	0	α_1	α_2	α_3	α_4
x	1	β_1	β_2	β_3	β_4
	2	0	0	1	0

(A_2)

			y		
		A	B	C	D
	0	α_1	α_2	α_3	α_4
x	1	α_1	α_2	α_3	α_4
	2	0	0	1	0

(A_3)

Figure 5.4
Specifications of the attraction parameters of one item from the ASCP Medical Laboratory Test for the alternative models A_1, A_2, and A_3.

In this example the hypothesized model (H) will be compared with three alternative models. The specifications of the attraction parameters of these models are depicted in Figure 5.4 for the first item. For the other items similar figures can be depicted. The first model (A_1) is a GSERE model, in which for each item all the attraction parameters are unequal to zero. Model A_2 is the same as Model A_1, but with the additional assumption that the correct alternative (i.e. alternative C) will be chosen if the subject is in the "Complete knowledge" state. Model A_3 has not only the same assumptions as Model A_2, but that of $\alpha_i=\beta_i$ (i=1,...,4) as well.

In Table 5.2 the likelihood-ratio test statistics, the Akaike's information criteria and Raftery's Bayesian information criterion for the four models are given. From the G^2 and AIC values we can conclude that the hypothesized model fits the data better then the alternative models. Furthermore, Model A_2 fits the data better than Model A_3, which means that there may be a significant difference between the attraction of the alternatives for a subject in the "Don't know" state and a subject in the "Partial knowledge" state. The better fit of Model A_2 in relation to Model A_1 may indicate that a subject in the "Complete knowledge" state would make no mistake in choosing the correct alternative.

Table 5.2
Likelihood-ratio test statistics, Akaike's information criteria and Raftery's Bayesian information criteria for four items from the ASCP Medical Laboratory Test

Model	Number of parameters	G^2 value	AIC value	BIC value
H	35	197.557	173.297	-86.736
A_1	53	328.733	291.996	-101.768
A_2	41	190.259	161.840	-142.770
A_3	29	235.102	215.001	-0.455
A_4	32	200.283	178.102	-59.642

In Table 5.3 the estimates of the attraction parameters for the alternatives of each item are presented for the hypothesized model. These results indicate that a subject in the "Partial knowledge" state is more likely to choose the correct alternatives to Items 1 and 4 than a subject in the "Don't know" state. In all probability a test constructor would never expect that a subject in the "Don't know" state is more likely to choose the correct alternative than a subject in the "Partial knowledge" state, as was the case for Item 2. The attraction parameter of the correct

Table 5.3
Attraction parameters for the alternatives of four items from the ASCP Medical Laboratory Test in the case of Model H

	Alternatives "Don't know" state				Alternatives "Partial knowledge" state			
Item	A	B	C	D	A	B	C	D
1	.204	.135	$\underline{.180}$	.481	*.000*	*.000*	$\underline{.676}$	.324
2	.027	.276	$\underline{.064}$	.633	.988	*.000*	$\underline{.012}$	*.000*
3	$\underline{.622}$	.163	.000	.215	$\underline{.502}$	.158	.340	*.000*
4	$\underline{.556}$	.120	.086	.238	$\underline{.828}$	.110	.063	*.000*

Note 1.: The correct alternatives are underlined.

Note 2.: Attraction parameters written in italics are prespecified.

alternative of Item 2 in the first state is five times larger than the associated parameter in the second state. However, in both states the probability of choosing the right alternative is very low. In the "Partial Knowledge" state alternative A is almost always chosen, whereas in the "Don't know" state alternative D is often chosen. In Item 2, therefore, we have the advantage that we can make a distinction between subjects who exactly knew the solution to the problem imposed by Item 2 and those who did not. The attraction parameters for the correct alternatives of Item 3 are approximately the same for both states.

Table 5.3 shows that some of the attraction parameters are smaller than .05. An interesting question for these cases is whether these attraction parameters are really unequal to zero or happen to have estimates unequal to zero by chance. Although it is bad practice to formulate an alternative model post hoc after looking at parameter estimates and test it on the same sample, we have tried to find an answer to the above question through the formulation of a fourth alternative model (A_4), which had the same assumptions as the hypothesized model (H). Another assumption was that all attraction parameters with estimated value smaller than .05 in the hypothesized model were equal to zero.

Table 5.4

Attraction parameters for the alternatives of four items from the ASCP Medical Laboratory Test in the case of Model A_4

	Alternatives "Don't know" state				Alternatives "Partial knowledge" state			
Item	A	B	C	D	A	B	C	D
1	.275	.182	.187	.355	*.000*	*.000*	.581	.419
2	*.000*	.286	.058	.656	1.000	*.000*	*.000*	*.000*
3	.623	.164	*.000*	.208	.597	.124	.279	*.000*
4	.477	.130	.112	.281	.919	.081	.000	*.000*

Note 1.: The correct alternatives are underlined.

Note 2.: Attraction parameters written in italics are prespecified.

This alternative model showed a slightly improved fit compared to the hypothesized model (see Table 5.2). From Table 5.3 we can also see that the alternative model fits the data better than all the other alternative models. In Table 5.4 the estimated attraction parameters for Model A_4 are given. It is clear that the estimated attraction parameters in the case of Model A_4 do not differ much from the estimated attraction parameters of the hypothesized model.

At this point we considered only the attraction parameters of the four items from the ASCP Medical Laboratory Test. In the remaining part of this section we will take a closer look at the difficulty parameters of the items. In this example the assumption was that the latent states were governed by the one-dimensional partial credit model. If the scoring weights for this model as depicted in Figure 5.1(b) are used and the latent trait index q is omitted, the one-dimensional version of Model (5.2) becomes

$$P(x|\theta) = \frac{\exp((\theta - \delta_j(x))x)}{\sum_z \exp((\theta - \delta_j(z))z)} = \frac{\exp\left(\sum_{g=1}^{x} (\theta - \psi_{jg})\right)}{\sum_z \exp\left(\sum_{g=1}^{z} (\theta - \psi_{jg})\right)} \tag{5.10}$$

in which $\psi_{jx} = x\,\delta_j(x) - (x-1)\,\delta_j(x-1)$ describes the difficulty of step x in item j, because each latent state may be seen as the result of a series of subsequent steps, each of which has to be taken. In Table 5.5 the values of the ψ_{jx} parameters for the four items of the ASCP Medical Laboratory Test in the case of the hypothesized model H are given.

Table 5.5
Estimated difficulty parameters of four items
from the ASCP Medical Laboratory Test for the case
of the hypothesized model H

Item	Step 1	Step 2
1	-2.15232	-0.64080
2	0.18140	-4.02920
3	-1.87987	-0.51876
4	-2.04992	1.20943

Table 5.5 shows that the difficulty of the steps changes positively for the items 1, 3 and 4, which means that it is more difficult to take the last step than the first step. On the other hand, for item 2 it is difficult to take the first step, but if the first step is reached the second step is very easy.

According to Verhelst and Verstralen (1991), two remarks have to be made. In the first place, as Molenaar (1983) showed, the parameter value of a particular step will depend on the parameter values for the other steps in the item. Therefore, the parameter value of a step cannot be interpreted as a measure of its difficulty. Secondly, it cannot be known in advance that the items allow for a sequential solution as assumed in the partial credit model.

Finally, some surprising results which are found during the analyses will be discussed. In the sample of the 3370 subjects no one had given a completely wrong answer to the four items. In the case of the hypothesized model, however, it was estimated that 68 subjects had no knowledge of the solution to the problems imposed by all four items. On the other hand, nearly 32% of the subjects gave a completely correct answer to all four items, whereas it was estimated that 2% really knew all the solutions to the problems.

5.9 DISCUSSION

In this chapter a loglinear item response theory (IRT) model with latent classes was proposed that related polytomously scored item responses to a multidimensional latent space. The proposed model is a generalization of the solution-error response-error (SERE) model (Kelderman 1988; Westers & Kelderman, 1992) to situations of polytomously scored latent states that may be explained by a multidimensional latent space. In this generalized SERE model (GSERE) a distinction was made between some well-defined latent states in which the subject has a certain amount of knowledge of the answer. The probability that the subject is in a certain state is assumed to be governed by the multi-dimensional polytomous latent trait model (MPLT). The relationship between the latent states and the observed answers is described through conditional probabilities.

Maximum likelihood estimates of the parameters of the GSERE model can be obtained by the IPF algorithm. The results by Westers and Kelderman (1992), however, indicate that the (G)SERE model is usable in practice only when the responses to a few items are studied. However, since the property of collapsibility is also valid for the GSERE model, the DBI-principle of Chapter 3 can be used for the estimation of larger sets of parameters in the GSERE model.

As pointed out in Westers and Kelderman (1992), an item can show DIF in two different ways. In the first place, an item shows DIF if equally able individuals from different subgroups have different probabilities of knowing the answer. Secondly, an item also shows DIF if the attractiveness of the alternatives varies from subgroup to subgroup. Just as in the case of the SERE model, the GSERE model can be extended with variables defining subgroups in order to study these two types of DIF. The GSERE model is, therefore, suitable for the examination of DIF in polytomous items through a combination of DIF for correct/incorrect responses and DIF in the alternatives.

Chapter 6

EPILOGUE

6.1 INTRODUCTION

The objective of this chapter is not only the summarization of the contents of the previous chapters, but also to make an inventory of those features of the differential item functioning (DIF) detection method based on the solution-error response-error (SERE) model that need further investigation.

Furthermore, we will indicate how the SERE model might be extended to increase its applicability for the examination of DIF.

6.2 SUMMARY

The subject of this dissertation is the examination of DIF with the use of loglinear Rasch models with latent classes. DIF is understood to describe the phenomenon that the probability of a correct response among equally able test takers is different for various racial, ethnic, or gender groups.

In Chapter 1 an overview was given of the DIF detection methods based on analysis of variance, on transformed item difficulties, on chi-square statistics, on item characteristic curves, on factor analysis, or on distractor response analysis. Although any of these methods can be used to detect biased items, they give little information about the reason why an item is biased. A reason for this omission is that existing DIF detection methods focus more or less on the observed responses and not on the process leading to the observed response. It was proposed, therefore, to use the SERE model of Kelderman (1988) for a more informative examination of DIF.

In Chapter 2 the SERE model was formulated and extended with variables for group membership. It was shown that with the SERE model we cannot only test whether an item shows DIF, but also whether DIF is caused by the difficulty of the item, the attractiveness of the alternatives, or both. However, the proposed method is not very practical for a large number of polytomous items, due to the fact that existing computer programs use a large amount of memory space, even for small sets of items.

In Chapter 3 a new estimation method was proposed. This new method of parameter estimation is based on the division of the whole item set into several subsets. This is possible because of the collapsability of the SERE model. It was shown that, depending on the number of items in each subset, the parameters of the SERE model can be estimated much more efficiently, both in terms of memory storage and processing time. However, some of the statistical efficiency of the estimators may be lost when the SERE model is collapsed. With the use of subsets of items, the parameters of the entire SERE model can only be obtained by simultaneous estimation of the parameters of the collapsed SERE models. It was shown that this can be achieved with the use of the pseudo-likelihood theory.

Chapter 4 dealt with the question whether DIF can be found with the proposed DIF detection method of this dissertation. This chapter also dealt with an examination of the new estimation method as introduced in Chapter 3. Therefore, in a simulation study, we examined how the values of the estimators differ from the true values. We also investigated whether this deviation is consistent in the sense that the differences tend to decrease when the number of items increases. Furthermore, we examined whether it is possible to detect an item which shows one type of DIF, but also shows the other type of DIF. Finally, we examined whether DIF can still be found if the number of items or the number of subjects is small. From this simulation study we concluded that with the SERE model a distinction can be made between both types of DIF and that it is possible to detect items which show both types of DIF.

In Chapter 5 a generalization of the SERE model applicable to polytomously scored latent states, that may be explained with a multidimensional latent space, was discussed. The critical difference between this model and the one-dimensional SERE model in Chapter 2 is that in the one-dimensional SERE model the probability that the subject is in a certain state is governed by the Rasch model, while in the generalized version of the SERE model this probability is governed by the multidimensional polytomous latent trait model of Kelderman and Rijkes (in press). The generalized SERE model can also be used for the examination of DIF and the parameters of this model can be estimated in a similar way as the parameters of the one-dimensional SERE model.

6.3 FUTURE RESEARCH

In this dissertation we showed how the SERE model can be used for the examination of DIF and how it can provide the test constructor or test user with more information about the nature of the bias factor. However, for the use of the SERE model in practice a new method for the estimation of the parameters of the SERE model needed to be developed. This new estimation method was based on the division of the entire item set into several subsets of items. By developing this new estimation method only one type of subsets of items was considered, namely a pair of items. One line of future research should be the construction of criteria for an optimum division of the entire item set into subsets of items. For instance, is a division into all possible pairs of items necessary or is a selection of these pairs sufficient? Moreover, will the estimation method be improved when the number of items in the subsets of items is three or more?

For the estimation of the parameters of the entire SERE model the pseudo-likelihood theory was used, in which we assumed that the loglikelihood of the entire SERE model is equal to the sum of the loglikelihoods of the collapsed SERE models over all subsets of items. The interesting question then arises whether the estimation method will be improved if, instead of the simple sum, a weighted sum of the loglikelihoods of the collapsed SERE models is used. Each weight in this summation may express, for example, the relative importance of the particular subset of items compared to the other subsets of items. A second question is whether possible optimum weights depend on the size of the subsets of items. Another line of future research should address this question.

Another issue with respect to the pseudo-likelihoods concerns the goodness of fit of the entire SERE model. The likelihood-ratio test statistics for each collapsed SERE model is chi-square distributed with degrees of freedom equal to the difference between the number of cells of the observed contingency table and the number of estimable parameters of the collapsed SERE model. However, is the (weighted) sum of these likelihoods-ratio test statistics over all subsets of items still chi-square distributed? Or can we develop other test statistics for the SERE model, like the Martin-Löf (1973) statistic, the statistics of van den Wollenberg (1979, 1982), or the statistics of Glas (1989)? Future research should also address the question whether the pseudo information criteria discussed in Chapter 3 have any practical or theoretical use for model selection.

Two further questions that should be considered in future research are the question of the applicability of the (generalized) SERE model and the question on which criteria a computerized DIF detection system should define an item as being biased. For instance, can the (generalized) SERE model be extended to models such as that of Mislevy and Verhelst (1990), in which different subjects are assumed to employ different strategies when responding to an item?

Should the computerized DIF detection system base his decision whether an item shows DIF on pseudo-likelihood or on another criterion? And what is the best strategy for detecting the biased items?

In Mislevy and Verhelst (1990) a model is presented for item responses when different subjects use different strategies, but only the responses and not the choice of strategy can be observed. In this model the assumption is that each subject belongs to one of a number of exhaustive and mutually exclusive classes, each with a unique item-solving strategy. Furthermore, the responses from all subjects in a given class are assumed to fit an IRT model. Finally, the assumption is that for each item its parameters under the IRT model for each strategy class can be related to known features of the item through psychological or substantive theory. As Mislevy and Verhelst stated, the main advantages of these multiple-strategy IRT models are that they provide a framework for testing alternative theories about cognitive processing, and the estimation of *how* subjects solve problems, in contrast to *how many* they solve. Such models could be used for diagnosis, remediation, and curriculum revision. With respect to this model, one line of future research could be an extension of the SERE model where the 'standard' IRT model is replaced by a multiple-strategy IRT model. In this way, more information can be obtained to solve the question why a subject is in a certain state.

The last line of future research which will be discussed in this chapter is the question on which criteria a computerized DIF detection system should define an item as biased and what the best strategy for the detection of the biased items is. Many authors, including Aitkin (1980) and Bishop, Fienberg and Holland (1975), have given guidelines for the selection of the best fitting models. These methods are mixed forms of forward and backward selection. For the selection of the biased items a similar approach can be used.

In the case of forward selection, the examination starts with a very restricted model, that is, a model in which no items show DIF. With successive defining items as being biased, we search for a model that has as few biased items as possible, but still has a good fit of the data. In the search for biased items, both the (consistent) pseudo Akaike's information criteria and the estimated values of the parameters of the model can be used. Backward selection procedures work the other way around: The examination starts with a model in which all the items show DIF in the latent response and DIF in the attraction parameters. With defining items as being unbiased, we search for a more economic model that still fits the data. For both procedures a careful study of the estimated values of the parameters of the model can indicate which items may or may not be assumed to show DIF. However, future research is needed to find the best selection procedure and the criteria on which a computerized DIF detection system should be based for each step in the procedure.

APPENDICES

A.1 THE ENGLISH VERSION OF ITEM 4 OF THE SECOND INTERNATIONAL MATHEMATICS STUDY

A quadrilateral MUST be a parallelogram if it has

A. one pair of adjacent sides equal
B. one pair of parallel sides
C. a diagonal as axis of symmetry
D. two adjacent angles equal
E. two pairs of parallel sides

A.2 THE COLLAPSED SOLUTION-ERROR RESPONSE-ERROR MODEL

In this appendix an instructional proof will be given that the SERE model is collapsible. Further, the same notation of Chapter 3 will be used. From elementary probability theory, it follows for the SERE model that

$$P(y_1,y_2) = \sum_{y_3} \dots \sum_{y_k} P(y_1,\dots,y_k) \equiv \sum_{y_3} \dots \sum_{y_k} P(y)$$

$$= \sum_{\mathbf{x}} \sum_{y_3} \dots \sum_{y_k} P(\mathbf{x})\, P(y|\mathbf{x}) = \sum_{\mathbf{x}} P(\mathbf{x}) \sum_{y_3} \dots \sum_{y_k} P(y|\mathbf{x})$$

$$= \sum_{\mathbf{x}} P(\mathbf{x})\, P(y_1|x_1)\, P(y_2|x_2) \sum_{y_3} \dots \sum_{y_k} P(y_3|x_3) \dots P(y_k|x_k)$$

in which P(**x**) is the marginal probability of the latent response vector **x**. From conditional probability calculus it follows that

$$P(y_1,y_2) = \sum_{\mathbf{x}} P(\mathbf{x})\; P(y_1|x_1)\, P(y_2|x_2)$$

$$= \sum_{x_1} \sum_{x_2} P(y_1|x_1)\, P(y_2|x_2) \sum_{x_3} \dots \sum_{x_k} P(\mathbf{x}).$$

With the use of the assumption of local independence in the Rasch model and elementary calculus it follows that

$$\sum_{x_3} \dots \sum_{x_k} P(\mathbf{x}) = \sum_{x_3} \dots \sum_{x_k} \int P(\mathbf{x}|\theta)\, dF(\theta)$$

$$= \int \sum_{x_3} \dots \sum_{x_k} \prod_{j=1}^{k} P(x_j|\theta)\, dF(\theta)$$

$$= \int P(x_1|\theta)\, P(x_2|\theta)\, dF(\theta).$$

The marginal probability of the observed responses y_1 and y_2 can then be written as

$$P(y_1,y_2) = \sum_{x_1} \sum_{x_2} P(y_1|x_1)\, P(y_2|x_2) \int P(x_1|\theta)\, P(x_2|\theta)\, dF(\theta)$$

$$= \sum_{x_1} \sum_{x_2} \Phi^{X_1Y_1}_{x_1y_1}\, \Phi^{X_2Y_2}_{x_2y_2}\, \Phi^{X_1}_{x_1}\, \Phi^{X_2}_{x_2}\, \Phi^{T12}_{z}$$

with $z=x_1+x_2$ and

$$\Phi^{T12}_{z} = \int \exp(z\theta)\, \{(1+\exp(\theta-\delta_1))\,(1+\exp(\theta-\delta_2))\}^{-1}\, dF(\theta).$$

A.3 LANPACO

LANPACO is a program for estimating the parameters of the loglinear Rasch model with latent classes. It uses the solution-error response-error (SERE) model of Kelderman (1988) and Westers and Kelderman (1992), and the estimation algorithm described in Chapter 3. The SERE model has been proven to be useful in the solution of practical psychometrics problems such as the examination of differential item functioning (DIF) or item bias (Westers & Kelderman, 1992) and the analysis of polytomously scored test items (Kelderman, 1988).

LANPACO includes procedures for testing whether an item exhibits DIF, estimating and testing the parameters of the SERE model, and graphical displaying of the results. For handling larger set of items, the estimation method and the goodness of fit test introduced in Chapter 3 is implemented in LANPACO. Finally, a simple and quick user interface has been added to the program.

The design of LANPACO allows interactive use of the program. If a model is disapproved, it is possible to change the model specifications and to reestimate the parameters of the model. This process can be repeated until the model is considered appropriate.

Since LANPACO is only a prototype, not all features of the SERE model or the estimation method are implemented in LANPACO. For instance, LANPACO will automatically select all possible pairs of items as subsets of items, but other selections of subsets of items are not possible. Further, in the program one can only specify an attraction parameter to be zero or not; other restrictions are not possible. All other features of the estimation method and the SERE model are implemented in LANPACO.

Finally, some technical notes are given. LANPACO was written in TURBO-PASCAL 6.0 under MS-DOS. The minimal configuration required is an AT compatible computer (80286 processor), but a 80386 or 80486 based machine is recommended. For the graphic user interface, LANPACO requires a video monitor with EGA or VGA graphics. However, VGA graphics will give better and clearer functions and tables. A coprocessor, if available, will increase the computing speed. At last, LANPACO has been developed at the Faculty of Educational Science and Technology of the University of Twente, the Netherlands.

A.4 TWO ITEMS OF THE AMERICAN SOCIETY OF CLINICAL PATHOLOGIST (ASCP) MICROBIOLOGY TEST

Item 11 Of the following bacteria, the most frequent cause of prosthetic heart valve infections occurring two to three months after surgery is:

a. Streptococcus pneumoniae
b. Streptococcus pyogenes
c. Staphylococcus aureus
d. Staphylococcus epidermidis

Item 20 A beta-hemolytic gram-positiv coccus was isolated from the cerebrospinal fluid of a 2-day-old infant with signs of meningtis. The isoltae grew on sheep blood agar under aerobic conditions and was resistant to a bacitracin disc. Which of the following should be performed for the presumptive identification of the organism?

a. oxidase production
b. catalase formation
c. CAMP test
d. esculin hydrolysis

A.5 THE COLLAPSED GENERALIZED SOLUTION-ERROR RESPONSE-ERROR MODEL

In this appendix an instructional proof will be given that Model 5.7 of Chapter 5 is collapsible. The same notation of Chapter 5 will be used. From elementary probability theory, it follows for the GSERE model that

$$P(y_1,y_2) = \sum_{y_3} \ldots \sum_{y_k} P(y_1,\ldots,y_k) \equiv \sum_{y_3} \ldots \sum_{y_k} P(\mathbf{y})$$

$$= \sum_{\mathbf{x}} \sum_{y_3} \ldots \sum_{y_k} P(\mathbf{x})\, P(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{x}} P(\mathbf{x}) \sum_{y_3} \ldots \sum_{y_k} P(\mathbf{y}|\mathbf{x})$$

$$= \sum_{\mathbf{x}} P(\mathbf{x})\, P(y_1|x_1)\, P(y_2|x_2) \sum_{y_3} \ldots \sum_{y_k} P(y_3|x_3) \ldots P(y_k|x_k)$$

in which P(**x**) is the marginal probability of the latent state vector **x**. From conditional probability calculus it follows that

$$P(y_1,y_2) = \sum_{\mathbf{x}} P(\mathbf{x})\, P(y_1|x_1)\, P(y_2|x_2)$$

$$= \sum_{x_1} \sum_{x_2} P(y_1|x_1)\, P(y_2|x_2) \sum_{x_3} \ldots \sum_{x_k} P(\mathbf{x})$$

and with the use of the assumption of local independence of the latent states x_j in the MPCM model and elementary calculus it follows that

$$\sum_{x_3}\dots\sum_{x_k} P(\mathbf{x}) = \sum_{x_3}\dots\sum_{x_k}\int P(\mathbf{x}|\theta)\,dF(\theta)$$

$$= \int \sum_{x_3}\dots\sum_{x_k}\prod_{j=1}^{k} P(x_j|\theta)\,dF(\theta)$$

$$= \int P(x_1|\theta)\,P(x_2|\theta)\,dF(\theta).$$

The marginal probability of the observed responses y_1 and y_2 can then be written as

$$P(y_1,y_2) = \sum_{x_1}\sum_{x_2} P(y_1|x_1)\,P(y_2|x_2)\int P(x_1|\theta)\,P(x_2|\theta)\,dF(\theta)$$

$$= \sum_{x_1}\sum_{x_2} \Phi^{X_1Y_1}_{x_1y_1}\,\Phi^{X_2Y_2}_{x_2y_2}\,\Phi^{X_1}_{x_1}\,\Phi^{X_2}_{x_2}\,\Phi^{T12}_{\mathbf{z}}$$

in which

$$\Phi^{T12}_{\mathbf{z}} = \int \exp\left(\sum_q z_q\theta_q\right)\{C(\theta,\delta_1)\,C(\theta,\delta_2)\}^{-1}\,dF(\theta),$$

$$C(\theta,\delta_j) = \sum_{\mathbf{z}} \exp\left\{\sum_{q=1}^{z}(\theta_q - \delta_{jq})\right\},$$

and $\mathbf{z}=(z_1,\dots,z_v)$ with $z_q = B_{1q}(x_1) + B_{2q}(x_2)$. Finally $P(y_1,y_2)$ is similar to Equation 5.7 of Chapter 5 except that here we consider two items and in Equation 5.7 k items. It may be clear that the collapsibility of the GSERE model is also valid, if the general MPLT model is used for the description of the dependence of the latent states on the latent traits.

REFERENCES

Ackerman, T.A., (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. ***Journal of Educational Measurement*, *29***, 67-91.

Aitkin, M. (1980). A note on the selection of log-linear models. ***Biometrics*, *36***, 173-178.

Akaike, H. (1977). On entropy maximizations principle. In P.R. Krisschnaiah, ***Application of statistics*** (p. 27-41). North Holland.

Akaike, H. (1987). Factor analysis and AIC. ***Psychometrika*, *52***, 317-332

Andersen, E.B. (1973). ***Conditional inference and models for measuring***. Doctoral dissertation. Copenhagen: Mentalhygiejnisk Forlag.

Anderson, T.W. (1954). On estimation of parameters in latent structure analyse. ***Psychometrika*, *19***, 1-10

Angoff, W.H. (1972). ***A technique for the investigation of cultural differences***. Paper presented at the annual meeting of the American Psychological Association, Honolulu.

Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.), ***Handbook of methods for detecting item bias***. Baltimore, MD:Johns Hopkins University Press.

Angoff, W.H., & Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude. ***Journal of Educational Measurement*, *10***, 95-105.

Arnold, B.C., & Strauss, D. (1988). ***Pseudo-likelihood estimation***. (Technical Report no. 164). Riverside, CA: Department of Statistics, University of California.

Baker, F.B. (1977). Advances in item analysis. ***Review of Educational Research*, *47***, 151-178.

Baker, F.B. (1987). Methodolgy review: Item parameters estimation under the one-, two-, and three-parameter logistic models. ***Applied Psycholical Measurement*, *11***, 111-141.

Baron, H. (1988). ***An investigation of differential item performance in Hebrew and Arabic versions of the University Psychometric entrance test***. Master thesis. Jerusalem: The Hebrew University of Jerusalem.

Bartholomew, D.J. (1987). ***Latent variable models and factor analysis***. London: Griffin.

Berk, R.A. (1982). ***Handbook of methods for detecting test bias***. Baltimore: The Johns Hopkins University Press.

Binet, A., & Simon, T. (1916). ***The development of intelligence in children***. Baltimore: Williams & Wilkins.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, ***Statistical theories of mental test scores***. Reading, Mass.: Addison-Wesley.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). ***Discrete multivariate analysis***. Cambridge, MA: MIT Press.

Bock, R.D. (1972). Estimating item parameters and latent proficiency when the responses are scored in two or more nominal categories. ***Psychometrika, 37***, 29-51.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. ***Psychometrika, 46***, 443-459.

Bock, R.D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. ***Journal of Educational Measurement, 25***, 275-285

Bozdogan, H. (1987). Model selection and Akaike's indormation criterion (AIC): The general theory and its analytical extensions. ***Psychometrika, 52***, 345-370

Cardall, C., & Coffmann, W.T. (1964). ***A method for comparing performance of different group on the items in a test***. (RM 64-61). Princeton, NJ: Educational Testing Service.

Cleary, T.A., (1968). Test bias: prediction of grades of negro and white students in integrated colleges. ***Journal of Educational Measurement, 5***, 115-124.

Cleary, T.A. & Hilton, T.L. (1968). An investigation into item bias. ***Educational and Psychological Measurement, 8***, 61-75.

Clogg, C.C. (1977). ***Unrestricted and restricted maximum likelihood latent structure analysis: A manual for users***. (Working paper no. 1977-09). University Park, PA.: Pennsylvania State University.

Clogg, C.C. (1981). Latent structure models of mobility. ***American Journal of Sociology, 86***, 836-868.

Clogg, C.C., & Goodman, L.A. (1985). Simultaneous latent structure analysis in several groups. In M.B. Tuma (Ed.), ***Sociological Methodology 1985*** (p. 81-110). San Francisco: Jossey-Bass Publishers.

Cramér, H. (1946). ***Mathematical methods of statistics***. Princeton: Princeton University Press.

Cressie, N., & Holland, P.W. (1983). Characterising the manifest probabilities of latent trait models. ***Psychometrika, 48***, 129-142.

Dayton, C.M., & Macready, G.B. (1980). A scaling model with response errors and intrinsically unscalable respondents. ***Psychometrika, 45***, 343-356.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM Algorithm. ***Journal of the Royal Statistical Society, Series B, 39***, 1-38.

Duncan, O.D. (1984). Rasch measurement: Further examples and discussion. In C.F. Turner, & E. Martin (Eds.), ***Surveying subjective phenomena, Vol. 2*** (p. 367-403). New York: Russell Sage Foundation.

Echternacht, G.A. (1974). A quick method for determining test bias. ***Educational and Psychological Measurement*, *34***, 271-280.

Eggen, T.J.H.M., Pelgrum, W.J., & Plomp, Tj. (1987). The implemented and attained mathematics curriculum: Some results of the second international mathematics study in the Netherlands. ***Studies in Educational Evaluation*, *13***, 119-135.

Eliason, S.R. (1988). ***The categorical data analysis system. Version 3.00A user's manual***. University Park, PA.: Pennsylvania State University, Department of Sociology.

Engelen, R.J.H. (1989). ***Parameter estimation in the logistic item response model***. Doctoral dissertation. Enschede: University of Twente, Faculty of Educational Science and Technology.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. ***Acta Psychologica*, *36***, 359-374.

Fischer, G.H. (1974). ***Einführung in die theorie psychologischer tests***. [Introduction into the theory of psychological tests]. Bern: Huber.

Fischer, G.H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. ***Psychometrika*, *46***, 59-77.

Fischer, G.H. (1987). Applying the principles of specific objectivity and of generalizability to the measurement of change. ***Psychometrika*, *52***, 565-587.

Formann, A.K. (1985). Constrained latent class models: Theory and applications. ***British Journal of Mathematical and Statistical Psychology*, *38***, 87-111.

Glas, C.A.W. (1989). ***Contribution to estimating and testing Rasch models***. Doctoral dissertation. Enschede: University of Twente, Faculty of Educational Science and Technology.

Goodman, L.A. (1974a). Exploratory latent structure analysis using both identifiable and unidentifiable models. ***Biometrika*, *61***, 215-231.

Goodman, L.A. (1974b). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I - a modified latent structure approach. ***The American Journal of Sociology*, *79***, 1179-1259.

Goodman, L.A. (1975). A new model for scaling response patterns: An application of the quasi-independence concept. ***Journal of the American Statistical Association*, *70***, 755-768.

Goodman, L.A. (1978). ***Analyzing qualitative/categorical data: Loglinear models and latent structure analysis***. London: Addison Wesley.

Green, D.R. (1976). ***Reducing bias in achievement tests***. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Green, B.F., Crone, C.R., & Folk, V.G. (1989). A method for studying differential distractor functioning. ***Journal of Educational Measurement*, *26***, 147-160.

Green, B.F., & Draper, J.F. (1972). ***Exploratory studies of bias in achievement tests***. Paper presented at the Annual Meeting of the American Psychological Association, Honolulu.

Guttman, L. (1950). The basis for scalogram analysis. In S.A. Stouffer et al. (Eds.), ***Measurement and prediction*** (p. 60-90). Princeton: Princeton University Press.

Haberman, S.J. (1979). ***Analysis of qualitative data. Vol. 2. New developments.*** New York: Academic Press.

Haberman, S.J. (1988). A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. ***Journal of the American Statistical Association*, *83***, 555-560.

Hagenaars, J. (1988). LCAG - Loglinear modelling with latent variables: A modified LISREL approach. In W.E. Saris, & I.N. Gallhofer (Eds.), ***Sociometric Research: Volume 2. Data Analysis*** (p. 111-130). London: MacMillan.

Hagenaars, J. (1990). ***Categorical longitudinal data: Log-linear panel, trend, and cohort analysis***. Newbury Park, CA: Sage Publications.

Hagenaars, J., & Luijkx, R. (1990). ***LCAG: latent class models and other loglinear models with latent variables***. (Working paper # 17^2). Tilburg: Tilburg University.

Hambleton, R.K., & Swaminathan, H. (1985). ***Item response theory; Principles and applications***. Boston: Kluwer-Nijhoff.

Holland, P.W., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel statistic. In H. Wainer & H. Braun (Eds.), ***Test validity*** (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hulin, C.L., Lissak, R.I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. ***Applied Psychological Measurement*, *6***, 249-260.

Hunter, J.E. (1975). ***A critical analysis of the use of item means and item-test correlation to determine the presence or absence of content bias in achievement test items***. Paper presented at the NIE Invitational Conference on Test Bias, Annapolis.

Ironson, G.H., & Craig, R. (1982). ***Item bias when amount of bias is varied and score differences between groups are present***. (Final Report NIE-G-81-0045). Tampa: University of South Florida.

Ironson, G.H., & Subkoviak, M.J. (1979). A comparison of several methods of assessing item bias. ***Journal of Educational Measurement*, *16***, 209-225.

Jensen, A.R. (1980). ***Bias in mental testing***. London: Methuen.

Kelderman, H. (1984). Loglinear Rasch model tests. ***Psychometrika***, ***49***, 223-245.

Kelderman, H. (1988). ***An IRT model for item responses that are subject to omission and/or intrusion errors***, (Research Report 88-16). Enschede: University of Twente, Faculty of Educational Science and Technology.

Kelderman, H. (1989). Item bias detection using loglinear IRT. ***Psychometrika***, ***54***, 681-697.

Kelderman, H. (1992). Computing maximum likelihood estimates of loglinear models from marginal sums with special attention to loglinear item response theory. ***Psychometrika***, ***57***, 437-450.

Kelderman, H., & Macready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. ***Journal of Educational Measurement***, ***27***, 307-327.

Kelderman, H., & Rijkes, C.P.M. (in press). Loglinear multidimensional IRT model for polytomously scored items. ***Psychometrika***.

Kelderman, H., & Steen, R. (1988). ***LOGIMO I: Loglinear item response theory modeling***. [Computer program]. Enschede: University of Twente, Faculty of Educational Science and Technology.

Koehler, K.J. (1977). ***Goodness-of-fit statistics for large sparse multinomials***. Unpublished Doctoral dissertation. Minneapolis: School of Statistics, University of Minnesota.

Koehler, K.J. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. ***Journal of the American Statistical Association***, ***81***, 483-493.

Kok, F. (1982). ***Het partijdige item***. [The biased item.] Amsterdam: Psychological Laboratory, University of Amsterdam.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), ***Latent trait and latent class models*** (pp. 263-274). New York: Plenum.

Kulback, S., & Leibler, R.A. (1951). On information and sufficiency. ***Annals of Mathematical Statistics***, ***22***, 79-86.

Lancaster, H.O. (1961). Significance tests in discrete distributions. ***Journal of the American Statistical Association***, ***56***, 223-234.

Langeheine, R., & van de Pol, F. (1990). A unifying framework for markov modeling in discrete space and discrete time. ***Sociological Methods and Research***, ***18***, 416-441.

Larnz, K. (1978). Small sample comparisons of exact levels for chi-square goodness-of-fit statistics. ***Journal of the American Statistical Association***, ***73***, 252-263.

Lazarsfeld, P.F., & Dudman, J. (1951). The general solution of the latent class case. In P.F. Lazarsfeld (Ed.), ***The use of mathematical models in the measurement of attitudes***. (RAND Research Memorandum no. 455).

Lazarsfeld, P.F., & Henry, N.W. (1968). ***Latent structure analysis***. Boston: Houghton-Miffin.

Lehmann, E.L. (1983). ***Theory of point estimation***. New York: John Wiley & Sons.

Linacre, J.M., & Wright, B.D. (1986). ***Item bias: Mantel-Haenszel and the Rasch model***. (Memorandum no. 39). Chicago: MESA, Psychometric Laboratory, Department of Education, University of Chicago.

Linn, R.L., & Harnisch, D.L. (1981). Interactions between item content and group membership on achievement test items. ***Journal of Educational Measurement, 18***, 109-118.

Linn, R.L., Levine, M.V., Hastings, C.N., & Wardrop, J.L. (1981). Item bias in a test of reading comprehension. ***Applied Psychological Measurement, 5***, 159-173.

Little, R.J.A., & Rubin, D.B. (1986). ***Statistical analysis with missing data***. New York: John Wiley & Sons.

Lord, F.M. (1975). ***Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters***. (Research Bulletin RB-75-33). Princeton, NJ.: Educational Testing Service.

Lord, F.M. (1977). A study of item bias using item characteristic curve theory. In Y.H. Poortinga (Ed.), ***Basic problems in cross-cultural psychology***. Amsterdam: Zwets and Zeitlinger.

Lord, F.M. (1980). ***Applications of item response theory to practical testing problems***. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F.M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. ***Psychometrika, 48***, 477-482

Lord, F.M., & Novick, M.R. (1968). ***Statistical theories of mental test scores***. Reading, MA: Addison-Wesley.

Lucassen, W.I., & Evers, A. (1984). ***Oorzaken en gevolgen van sexe-partijdigheid in de Differentiële Aanleg Testserie DAT '83***. (Causes and consequences of sex-bias in the Differential Aptitude Test series DAT 1983.) Paper presented at the Congress of Dutch Psychologists, Ede, The Netherlands.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. ***Journal of the National Cancer Institute, 39***, 719-748.

Macready, G.B., & Dayton, C.M. (1980). The nature and use of state mastery models. ***Applied Psychological Measurement, 4***, 493-516.

Marascuillo, L.A., & Slaughter, R.E. (1981). Statistical procedures for identifying possible sources of item bias based on X^2 statistics. ***Journal of Educational Measurement, 18***, 229-248.

Martin-Löf, P. (1973). ***Statistika Modeller. Anteckningar från seminarier Lasåret 1969-1970 utarbetade av Rolf Sunberg obetydligt ändrat nytryk***. Stockholm: Institutet för Försäkringsmatematik och Matematisk Statistik vid Stockholms Universitet.

Masters, G.N. (1982). A Rasch model for partial credit scoring. ***Psychometrika*, *47***, 49-72.

Maxwell, A.E. (1959). Maximum likelihood estimates of item parameters using the logistic function. ***Psychometrika*, *24***, 221-227.

McHugh, R.B. (1956). Efficient estimation and local identification in latent class analysis. ***Psychometrika*, *21***, 331-347.

Mellenbergh, G.J. (1982). Contingency table methods for assessing item bias. ***Journal of Educational Statistics*, *7***, 105-118.

Mellenbergh, G.J. (1985). Vraag-onzuiverheid: definitie, detectie en onderzoek. (Item bias: Dutch research on its definition, detection and explanation.) ***Nederlands Tijdschrift voor de Psychologie*, *40***, 425-435.

Mellenbergh, G.J. (1989). Item bias and item response theory. In R.K. Hambleton (Ed.), Applications of item response theory [Special issue]. ***International Journal of Educational Research*, *13***, 127-143.

Mellenbergh, G.J., & Kok, F.G. (in press). Finding the biasing trait(s). In S.H. Irvine, S. Newstead, & P. Dann (Eds.). ***Computer-based human assessment, Part 3, Identifying group and individual patterns of response to tests***. Boston, MA: Academic Publishers.

Mellenbergh, G.J., & Vijn, P. (1981). The Rasch model as a loglinear model. ***Applied Psychological Measurement*, *5***, 369-376.

Merz, W.R. (1973). ***Factor analysis as a technique in analyzing test bias***. Paper presented at the Annual Meeting of the California Educational Research Association, Los Angeles.

Merz, W.R. (1976). ***Estimating bias in test items utilizing principal components analysis and the general linear solution***. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Mislevy, R.J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solutions strategies. ***Psychometrika*, *55***, 195-216.

Molenaar, I.W. (1983). Some improved diagnostics for the Rasch model. ***Psychometrika*, *48***, 49-72.

Mooijaart, A. (1978). ***Latent structure models***. Doctoral dissertation. Leiden: University of Leiden.

Mooijaart, A., & van der Heijden, P.G.M. (1992). The EM algorithm for latent class analysis with equality constraints. ***Psychometrika*, *57***, 261-269.

Muraki, E., & Engelhard, G. (1989). ***Examining differential item funtioning with BIMAIN***. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. ***Journal of Educational Statistics*, *10***, 133-142.

Neyman, J. (1949). Contribution to the theory of the X^2-test. In ***Proceedings of the Berkeley Symposium on Mathematical Statistics and Probablity***. Berkeley: University of California Press.

Osterlind, S.J. (1983). ***Test item bias***. Beverly Hills: Sage.

Paulson, J.A. (1986). ***Estimation of parameters in latent class models with constraints on the parameters***. (Technical Report ONR86-1). Portland: Psychology Department, Portland State University.

Petersen, N.S. (1977). ***Bias in the selection rule: Bias in the test***. Paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands.

Raftery, A.E. (1986a). Choosing models for cross-classifications. ***American Sociological Review, 51***, 145-146.

Raftery, A.E. (1986b). A note on Bayes factors for log-linear contingency table methods with vague prior information. ***Journal of the Royal Statistical Society. Series B, 48***, 249-250.

Raju, N.S. (1988). The area between two items characteristic curves. ***Psychometrika, 53***, 495-502.

Raju, N.S., Bode, R.K., & Larsen, V.S. (1987). ***An empirical assessment of the Mantel-Haenszel statistic for studying differential item functioning***. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Rao, C.R. (1973). ***Linear statistical inference and its applications***. New York: Wiley.

Rasch, G. (1960/1980). ***Probabilistic models for some intelligence and attainment tests***. Copenhagen: Paedagogiske Institut. (Expanded edition, Chicago: The University of Chicago Press, 1980)

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. ***Applied Psychological Measurement, 3***, 271-282.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. ***The British Journal for Mathematical and Statistical Psychology, 44***, 75-92.

Rost, J., & von Davier, M. (1992). ***MIRA: A PC-program for the mixed Rasch model. User manual***. Kiel: Institute for Science Education.

Rubin, D.B. (1991). EM and beyond. ***Psychometrika, 56***, 241-254.

Rudner, L.M., & Convey, J.J. (1978). ***An evaluation of select approaches for biased item identification***. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto.

Rudner, L.M., Getson, P.R., & Knight, D.L. (1980a). Biased item detection techniques. ***Journal of Educational Statistics, 5***, 213-233.

Rudner, L.M., Getson, P.R., & Knight, D.L. (1980b). A Monte Carlo comparison of seven biased item detection techniques. ***Journal of Educational Measurement*, *17***, 1-10.

Scheiblechner, H. (1972). Das lernen und lösen komplexer denkaufgaben. [Learning and solving complex cognitive problems.] ***Zeitschrift für experimentelle und angewandte psychologie*, *19***, 476-506.

Scheuneman, J.D. (1979). A method of assessing bias in test items. ***Journal of Educational Measurement*, *16***, 143-152.

Scheuneman, J.D. (1987). An experimental, exploratory study of causes of bias in test items. ***Journal of Educational Measurement***, 24, 97-118.

Schmitt, A.P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. ***Journal of Educational Measurement*, *25***, 1-13.

Schwarz, G. (1978). Estimating the dimensions of a model. ***Annals of Statistics*, *6***, 461-464.

Sclove, S.L. (1987). Application of model-selection criteria to some problems in multivariate analysis. ***Psychometrika*, *52***, 333-343.

Shealy, R., & Stout, W. (1991). ***A procedure to detect test bias present simultaneously in several items***. (Technical Report No. 91-3-ONR). Champaign: University of Illinois.

Shepard, L.A., Camilli, G., & Averill, M. (1981). Comparisons of procedures for detecting test-item bias with both internal and external ability criteria. ***Journal of Educational Statistics*, *6***, 317-377.

Shepard, L.A., Camilli, G., & Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. ***Journal of Educational Statistics*, *9***, 93-128.

Shepard, L.A., Camilli, G., & Williams, D.M. (1985). Validity of approximation techniques for detecting item bias. ***Journal of Educational Measurement*, *22***, 77-105

Smith, J.K., & Camilli, G.P. (1988). ***Recognizable subsets of examinees who disproportionately contribute to differential item functioning***. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans

Stricker, L.J. (1981). ***A new index of differential subgroup performance: Applications to the GRE Aptitude Test***. (GRE Research Report 78-7) Princeton, NJ: Educational Testing Service.

Subkoviak, M.J., Mack, J.S., Ironson, G.H., & Craig, R.D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. ***Journal of Educational Measurement***, 21, 49-58.

Swaminathan, H., & Gifford, J.A. (1983). Estimation of parameters in the three-parameter logistic model. In D.J. Weiss (Ed.), ***New horizons in testing: Latent trait test theory and computerized adaptive testing*** (p. 13-30). New York: Academic Press.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one parameter logistic model. ***Psychometrika*, *47***, 175-186.

Thissen, D., & Mooney, J.A. (1989). Loglinear item response models, with applications to data from social surveys. In C.C. Clogg (Ed.), ***Sociological Methodology 1989*** (p. 299-330). San Francisco: Jossey-Bass Publishers.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. ***Psychometrika*, *49***, 501-519

Thissen, D., Steinberg, L., & Fitzpatrick, A.R. (1989). Multiple choice models: The distractors are also part of the item. ***Journal of Educational Measurement*, *26***, 161-176.

Thissen, D., Steinberg, L., & Gerard, M. (1986). Beyond group-mean differences: The concept of item bias. ***Psychological Bulletin*, *99***, 118-128

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland, & H. Wainer (Eds.), ***Differential item functioning: Theory and practice***. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory, ***Psychometrika*, *47***, 397-412.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. ***Scandinavian Journal of Statistics*, *9***, 23-30.

Van de Pol, F.J., & Langeheine, R. (1990). Mixed markov latent class models. ***Sociological Methodology***, ,213-247.

Van de Pol, F.J., Langeheine, R., & de Jong, W. (1989). ***PANMARK user manual: Panel analysis using Markov chains, version 1.5***. Voorburg: Netherlands Central Bureau of Statistics.

Van den Wollenberg, A.L. (1979). ***The Rasch model and the time limit tests***. Doctoral dissertation. Nijmegen: Catholic University of Nijmegen.

Van den Wollenberg, A.L. (1982). Two new test statistics for the Rasch model. ***Psychometrika*, *47***, 123-140.

Van der Flier, H. (1982). ***Some applications of an iterative method to detect biased items***. Paper presented at the Sixth International Conference of the International Association of Cross-Cultural Psychology, Aberdeen, Scotland.

Veale, J.R., & Foreman, D.I. (1983). Assessing cultural bias using foil response data: cultural variation. ***Journal of Educational Measurement*, *20***, 249-258.

Verhelst, N.D., Glas, C.A.W., & van der Sluis, A. (1984). Estimation problems in the Rasch model: The basic symmetric functions. ***Computational Statistics Quarterly*, *1***, 245-262.

Verhelst, N.D., & Veldhuijzen, H.H. (1991). ***A new algorithm for computing elementary symmetric functions and their first and second derivatives***. (Measurement and Research Department Reports 91-1). Arnhem: Cito.

Verhelst, N., & Verstralen, H.H.F.M. (1991). ***The partial credit model with non-sequential solution strategies***. (Measurement and Research Department Reports 91-5). Arnhem: Cito

Westers, P., & Kelderman, H. (1992). Examining differential item functioning due to item difficulty and alternative attractiveness. ***Psychometrika*, *57***, 107-118.

Westers, P., & van der Sar, D. (1993). LANPACO: Estimating the parameters of loglinear latent class analysis models using pairs of items. [Computer program]. Enschede: University of Twente, Faculty of Educational Science and Technology.

Wingersky, M.S., & Lord, F.M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. ***Applied Psychological Measurement*, *8***, 347-364.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. ***Journal of Educational Measurement*, *14***, 97-116.

Wright, B.D., Mead, R.J., & Draba, R. (1975). ***Detecting and correcting test item bias with a logistic response model*** (RM 22), Chicago: Statistical Laboratory, Department of Education, University of Chicago.

Wu, C.F.J. (1983). A note on the uniqueness of minimum rank solutions in factor analysis. ***Psychometrika*, *46***, 109-110.

Yamamoto, K. (1987). ***A model that combines IRT and latent class models***. Illinois: University of Illinois at Urbana-Champaign.

Yamamoto, K. (1988). ***Hybrid model of IRT and latent class models***. Princeton, NJ: Educational Testing Service.

Zwick, E. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? ***Journal of Educational Statistics*, *15***, 185-197.

NEDERLANDSE SAMENVATTING
(DUTCH SUMMARY)

Op het terrein van het onderwijskundig en/of psychologisch meten is de formulering van de vragen in een test een van de facetten die beschouwd kunnen worden. Naast een paar algemene zaken als vormgeving, volgorde van de vragen en het taalgebruik, moet ervoor worden gezorgd dat bij het opstellen van de vragen allerlei bronnen van misleiding worden vermeden. Een belangrijk punt hierbij is, dat de (formulering van de) vraag aanleiding kan geven tot systematisch verschillende antwoorden van respondenten met gelijke vaardigheid. Onder vaardigheid wordt hier verstaan de vaardigheid om de kennis omtrent het te toetsen onderwerp toe te passen. Binnen de testtheorie wordt het geven van systematisch verschillende antwoorden door respondenten met dezelfde vaardigheid ook wel aangegeven met het begrip vraagpartijdigheid. In het voorliggende proefschrift zal er op dit begrip verder worden ingegaan en zal er een methode worden gepresenteerd om te toetsen of een bepaalde vraag partijdig is ten opzichte van bepaalde categorieën van respondenten. Daarbij worden alleen de uit meerkeuzevragen bestaande vragenlijsten beschouwd.

Het begrip vraagpartijdigheid, dat ook wel wordt aangeduid met de Engelse begrippen "item bias" of "differential item functioning", kan men op verschillende manieren definiëren. In het proefschrift zal de term "differential item functioning" (DIF) worden gebruikt. Een item (vraag) vertoont DIF als respondenten met gelijke vaardigheid een ongelijke kans hebben om het item correct te beantwoorden. Met andere woorden een item vertoont DIF als respondenten uit de ene groep (de "Focal" groep) niet dezelfde kans heeft om het item correct te beantwoorden als respondenten met dezelfde vaardigheid uit een andere groep (de "Reference" groep). De definitie van DIF houdt in dat verschillen tussen de scores op de toets (testscores) niet zonder meer als verschil in vaardigheid kunnen worden geïnterpreteerd.

Om te bepalen of een item DIF vertoont zijn er in de afgelopen jaren vele methoden ontwikkeld, waarvan in hoofdstuk 1 de belangrijkste worden besproken. Ruwweg kunnen deze DIF detectie methoden worden onderverdeeld in twee groepen, namelijk methoden die rekenschap houden met de vaardigheid van de respondent en de (vroegere) methoden die dat niet doen. Binnen de tweede groep vallen methoden die gebaseerd zijn op de variantie analyse of getransformeerde item moeilijkheden. Methoden die gebaseerd zijn op chi-kwadraat statistieken, factor analyse, analyse van de onjuiste antwoorden (alternatieven) of een item karakteristieke kromme behoren tot de eerste groep. Over het algemeen hebben methoden uit de eerste groep de voorkeur, omdat zij beter in staat zijn DIF te onderscheiden van verschillen in vaardigheden van de groepen respondenten.

Alhoewel elk van de methoden uit de eerste groep gebruikt kunnen worden voor de detectie van DIF, geven ze weinig informatie over de aard van de factoren die DIF hebben veroorzaakt. Een reden hiervoor is dat bestaande DIF detectie methoden over het algemeen gericht zijn op de geobserveerde antwoorden en niet op het proces dat heeft geleid tot de geobserveerde antwoorden. Bij meerkeuzevragen zal voordat een antwoord gegeven kan worden, eerst het probleem van de vraag onderkend en opgelost moeten worden. Op ieder niveau van dit proces kan er DIF optreden. De kans dat een probleem met succes onderkend en opgelost wordt hangt af van de moeilijkheidsgraad van het probleem. Deze moeilijkheidsgraad kan verschillend zijn voor verschillende groepen respondenten met dezelfde vaardigheid, hetgeen zou betekenen dat het item DIF vertoont.

Afhankelijk van de vraag of het probleem opgelost kon worden of niet, zal de respondent één van de antwoordcategorieën moeten kiezen. De keuze van een antwoord hangt daarbij af van de aantrekkelijkheid van het antwoord. Als nu de aantrekkelijkheid van de antwoordcategorieën voor verschillende groepen respondenten met dezelfde vaardigheid verschillen, dan vertoont het item ook DIF. In het proefschrift wordt een model bestudeerd dat het mogelijk maakt om beide typen van DIF gecombineerd en tegelijkertijd te bekijken, namelijk het solution-error response-error (SERE) model.

Zoals de naam van het model al suggereert bestaat het SERE model uit twee delen. Binnen het SERE model wordt ten eerste onderscheid gemaakt tussen twee toestanden waarin een respondent zich kan bevinden; de respondent weet de volledige oplossing van het probleem of de respondent kent die oplossing niet. Er wordt vervolgens aangenomen dat de kans dat een respondent de oplossing volledig weet, dus dat de respondent zich in de eerste toestand bevindt, beschreven wordt door het dichotome Rasch model. Dit model wordt bepaald door een logistische functie van het verschil tussen de moeilijkheidsgraad van het item en de vaardigheid van de respondent.

Het tweede deel van het SERE model beschrijft de uiteindelijke keuze van de respondent voor een bepaald antwoordcategorie. Aangenomen wordt, dat als de respondent de oplossing niet weet, de respondent het voor hem/haar dan meest aantrekkelijke antwoordcategorie als mogelijk juiste antwoord kiest. De relatie tussen de latente response van een respondent en zijn/haar uiteindelijke keuze van een antwoordcategorie, wordt in het SERE model gedefineerd als een conditionele kans. Relatief hoge waarden van deze conditionele kans geven aan dat de bijbehorende antwoordcategorie relatief zeer aantrekkelijk is, gegeven de toestand waarin de respondent zich bevindt. In het voorliggende proefschrift wordt verder aangenomen dat indien de respondent de oplossing wel weet, de respondent altijd de juiste antwoordcategorie kiest.

In hoofdstuk 2 wordt het SERE model verder besproken. Tevens wordt in dit hoofdstuk aangegeven dat met het SERE model het niet alleen mogelijk is om te onderzoeken of een item DIF vertoont, maar ook of DIF veroorzaakt wordt door de moeilijkheidsgraad van het item en/of door de aantrekkelijkheid van de antwoordcategorieën. Ook wordt aangegeven hoe met behulp van bestaande programmatuur de parameters van het model, dat wil zeggen de moeilijkheidsgraad van de items, de vaardigheden van de respondenten en de aantrekkelijkheden van de antwoordcategorieën, geschat kunnen worden.

Als het aantal items in een test te groot wordt, dan is het in de praktijk onmogelijk om met de bestaande programmatuur de parameters te schatten. Om dit probleem op te lossen wordt in hoofdstuk 3 een alternatieve schattingsmethode aangedragen. Hierin wordt de verzameling items verdeeld in een aantal deels overlappende deelverzamelingen. Door nu de parameters van iedere deelverzameling simultaan te schatten, kan men de parameters van de gehele test efficiënt schatten. Een bijkomend voordeel is dat de antwoorden van respondenten die alleen een gedeelte van de test hebben ingevuld, toch gebruikt kunnen worden bij het schatten van de moeilijkheden van de items en de aantrekkelijkheden van de antwoordcategorieën.

Om te bekijken of deze nieuwe schattingsmethode en de op het SERE model gebaseerde DIF detectie methode betrouwbare resultaten opleveren, wordt in hoofdstuk 4 een simulatie studie uitgevoerd. Vragen die bij deze studie centraal staan zijn: (1) kan DIF nog steeds worden aangetoond als het aantal items of het aantal respondenten klein is?; (2) in hoeverre verschillen de geschatte waarden van de parameters van de oorspronkelijke waarden van de parameters in de gesimuleerde data?; (3) is dit verschil consistent in de zin van dat de verschillen de neiging hebben om kleiner te worden als het aantal respondenten toeneemt? Tenslotte worden in dit hoofdstuk de minimum condities bestudeerd waaronder het SERE model en de daarop gebaseerde DIF detectie methode nog praktisch bruikbaar zijn.

Tot zover is ervan uitgegaan dat er maar één vaardigheid wordt getoetst (bijvoorbeeld rekenen of de Engelse taal) en dat er maar twee toestanden zijn: de respondent weet het antwoord of de respondent weet het niet. In de praktijk zijn er echter vele situaties waarbij van de respondenten meerdere vaardigheden worden verwacht. Zo worden bij de vraag "Wat is de wortel van zestien plus negen?" drie operaties verwacht. Ten eerste zal de respondent de vraag moeten vertalen in een wiskundige formule, om vervolgens de wortel van 16 uit te rekenen en als laatste de optelling uit te voeren. Daarnaast zijn er situaties waarbij de respondent een gedeelte van de oplossing van het probleem weet, maar niet in staat is om het probleem in zijn geheel op te lossen. Het is mogelijk dat een respondent het bovenstaande item wel kan vertalen in een wiskundige formule en ook nog de optelling uit kan voeren, maar niet weet hoe hij/zij de wortel uit een getal moet berekenen. Deze situatie is dus een voorbeeld waarbij een respondent zich in drie toestanden kan bevinden: de respondent weet de oplossing niet, de respondent weet een gedeelte van de oplossing of de respondent weet de volledige oplossing. In hoofdstuk 5 wordt het SERE model gegeneraliseerd tot dergelijke situaties. Analoog aan het SERE model, kan ook de gegeneraliseerde versie van het SERE model gebruikt worden als DIF detectie methode. De parameters van het gegeneraliseerde SERE model kunnen eveneens geschat worden met de methode die in hoofdstuk 3 geïntroduceerd is.

Het proefschrift eindigt in hoofdstuk 6 met een overzicht van nog nader te bestuderen kenmerken en/of eigenschappen van de op het (gegeneraliseerde) SERE model gebaseerde DIF detectie methode en de daarbij ontwikkelde nieuwe schattingsmethode. Te denken valt hierbij onder meer aan de optimale opdeling van de verzameling items in deelverzamelingen of aan de optimale methode om de items die DIF vertonen te selecteren. Verder wordt aangegeven hoe het (gegeneraliseerde) SERE model nog verder uitgebreid zou kunnen worden zodat de bruikbaarheid voor het bestuderen van DIF vergroot wordt.

CURRICULUM VITAE

The writer of this dissertation was born on 20 August 1960 in the city of Groningen. After he graduated at the Minister Cort van de Linden MAVO school in 1977 and at the Rijksscholengemeenschap Kamerlingh Onnes HAVO school in Groningen in 1979, he did a teacher training course in Mathematics and Economics at the Ubbo Emmius Institute for Secondary Teacher Training, also in Groningen. In 1985 he enrolled as a Mathematics student at the State University of Groningen with Statistics as his main subject. He got his Masters degree in March 1988 with the Master's thesis 'Does Telepathy exist? Data of Heymans et al. reanalyzed'. After a short period of unemployment he was appointed research assistant (AIO) to the research project titled 'Item bias detection using the loglinear Rasch model with latent classes' at the Faculty of Educational Science and Technology of the University of Twente. This research was done under the auspices of the Interuniversity Graduate School of Psychometrics and Sociometrics. Since August 1992 he has been working as an assistant professor at the Center for Biostatistics at the University of Utrecht.

CURRICULUM VITAE

De schrijver van dit proefschrift werd op 20 augustus 1960 te Groningen geboren. Na het in 1977 behalen van zijn MAVO diploma aan de Minister Cort van der Linden school te Groningen en het in 1979 behalen van zijn HAVO diploma aan de Rijksscholengemeenschap Kamerlingh Onnes te Groningen, volgde hij een opleiding voor docent Wiskunde en Economie aan het Instituut Lerarenopleiding Voortgezet Onderwijs Ubbo Emmius eveneens te Groningen. In 1985 begon hij aan de Rijksuniversiteit Groningen met de studie Wiskunde met als afstudeerrichting Statistiek. Het doctoraalexamen werd afgelegd in maart 1988 met de doctoraalscriptie getiteld "Bestaat telepathie? Data van Heymans e.a. opnieuw geanalyseerd". Na een korte periode van werkloosheid werd hij in augustus 1988 aangesteld als assistent in opleiding (AIO) op het onderzoeksproject getiteld "Item bias detectie met behulp van het loglineaire Rasch model met latente klassen" bij de Faculteit der Toegepaste Onderwijskunde van de Universiteit Twente. Dit onderzoek werd uitgevoerd onder auspiciën van het Interuniversitair Onderzoeksinstituut voor Psychometrie en Sociometrie. Sinds augustus 1992 is hij werkzaam als universitair docent bij het Centrum voor Biostatistiek aan de Universiteit Utrecht.